The Design of VoIP Systems With High Perceptual Conversational Quality

Benjamin W. Wah and Batu Sat
Department of Electrical and Computer Engineering and the Coordinated Science Laboratory,
University of Illinois, Urbana-Champaign, Urbana, IL 61801, USA
Email: {wah, batusat}@illinois.edu

***Abstract*— This paper describes our work on real-time two-party and multi-party VoIP (voice-over-IP) systems that can achieve high perceptual conversational quality. It focuses on the fundamental understanding of conversational quality and its trade-offs among the design of speech codecs and strategies for network control, playout scheduling, and loss concealments. We have studied three key aspects that address the limitations of existing work and improve the perceptual quality of VoIP systems. Firstly, we have developed a statistical approach based on just-noticeable difference (JND) to significantly reduce the large number of subjective tests, as well as a classification method to automatically learn and generalize the results to unseen conditions. Using network and conversational conditions measured at run time, the classifier learned helps adjust the control algorithms in achieving high perceptual conversational quality. Secondly, we have designed a cross-layer speech codec to interface with the loss-concealment and playout scheduling algorithms in the packet-stream layer in order to be more robust and effective against packet losses. Thirdly, we have developed a distributed algorithm for equalizing mutual silences and an overlay network for multi-party VoIP systems. The approach leads to multi-party conversations with high listening only speech quality and balanced mutual silences.**

I. INTRODUCTION

This paper summarizes our results on real-time two-party and multi-party VoIP (*voice-over-IP*) systems that can achieve high perceptual conversational quality. It focuses on the fundamental understanding of conversational quality and its trade-offs among the design of speech codecs and strategies for network control and loss concealments. An important aspect of this research is on the development of new methods for reducing the large number of subjective tests and for automated learning and generalization of the results of subjective evaluations. Since the network delays in VoIP can be long and time varying, its design is different from those for PSTN (*public switched telephone network*) with short and consistent delays [1], [2].

Figure 1 outlines the components in the design of a VoIP system. The first component on *conversational quality* entails the study of human conversational behavior, modeling conversational dynamics, and identifying user-perceptible attributes that affect quality. Its study also includes the design of off-line subjective tests and algorithms for learning the test results. Next, the study of *network* and *conversational environments* entails the identification of objective metrics for characterizing network and conversational conditions and the dissemination of this information at run time. The design of *loss-concealment* (LC) and *playout scheduling* (POS) strategies in the *packet-stream layer* involves delay-quality trade-offs that optimize user-perceptible attributes. The *network-control layer* provides support for network transport and admissions control in multi-party VoIP. Lastly, the design of the *speech codec* and its LC and compression capabilities must take into account its interactions with the LC and POS strategies in the packet-stream layer.

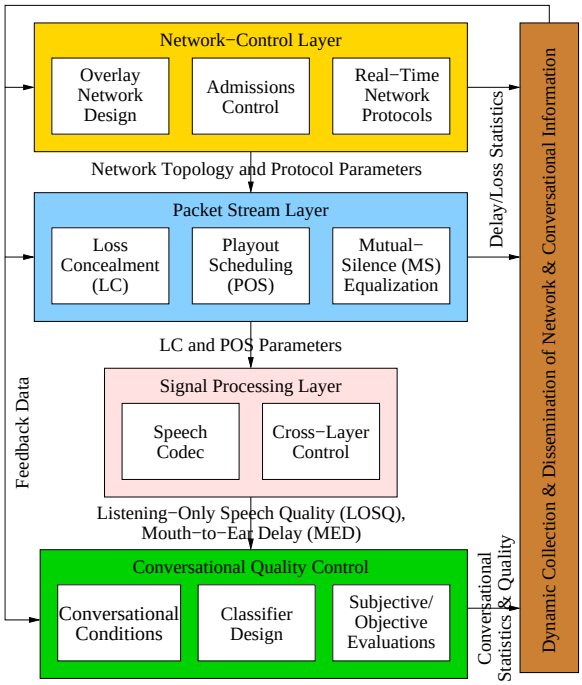

Figure 1. Layers in the architecture of a VoIP system.

Effects of delays on conversations. In a two-party conversation, each participant takes turns in speaking and listening [3]–[5], and both perceive a silence period (called *mutual silence* or *MS*) when the conversation switches from one party to another. Hence, a conversation consists of alternating speech segments and silence periods.

In a face-to-face setting, both participants have a common reality of the conversation: one speech segment

This work was supported by National Science Foundation grant CNS 08-41336.

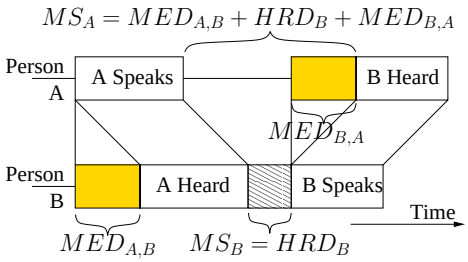

Figure 2. Asymmetric mutual silences

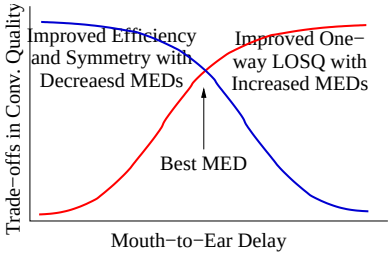

Figure 3. Trade-off considerations

is separated from another by a silence period that is identically perceived by both. However, when the same conversation is conducted over the Internet, the participants' perception of the conversation is different due to delays, jitters, and losses incurred on the speech segments during their transmission [6], [7].

Richards [8] has identified three factors that influence the quality of service in telephone systems: difficulty in listening to one-way speech, difficulty in talking, and difficulty in conversing during turn-taking. Hence, we evaluate the quality of a conversation over a network connection by the quality of the one-way speech segments received (the *listening-only speech quality* or *LOSQ*) and that of the interactions [6]; the latter is measured by the delay incurred from the mouth of the speaker to the ear of the listener (the *mouth-to-ear delay* or *MED*) [8].

When a connection has delays, the MSs perceived by a participant consist of alternating short and long silence periods between turns [6]. This asymmetry is caused by the fact that after A speaks, the MS experienced by A (MS_A in Figure 2) consists of the time for A's speech to travel to B ($MED_{A,B}$), the time for B to construct a response (*human response delay* or HRD_B), and the time for B's response to travel to A ($MED_{B,A}$). In contrast, after B hears the speech from A, the MS experienced by B is only governed by his/her HRD ($MS_B = HRD_B$). This asymmetry leads to a perception that each user is responding slowly to the other, and consequently results in degraded efficiency and perceptual quality [6].

Conversational quality cannot be improved by simultaneously improving LOSQ and reducing MED. A longer MED will improve LOSQ because segments will have a higher chance to be received, but will worsen the symmetry of MSs. Figure 3 shows the delay-quality trade-off and a suitable MED with the best quality. This trade-off also depends on the turn-switching frequency [6], [9] and on changes in network and conversational conditions [10]. It has been shown that long MEDs can cause double-talks and interruptions even when MED is constant [8], [11].

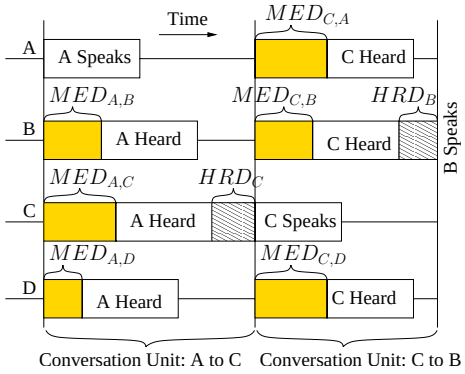

Figure 4. A 4-party conversation.

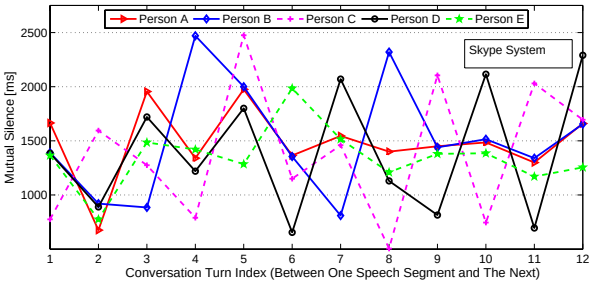

Figure 5. MSs in a simulated Skype conversation.

The perceived effects of delays in multi-party VoIP is more complex than those in two-party VoIP because there may be large disparities in network conditions across the participants [12]. In the multi-party case, the conversational quality depends on the LOSQ and the latency of the one-way speech from each speaker, as well as the symmetry of the conversation among the participants. Hence, each listener may have a slightly different perception of the same conversation. Figure 4 depicts two *conversation units* (CU) in a 4-party conversation. Each listener experiences different asymmetric MSs across the speakers: some appear to be more distant than others, or some respond slower than others.

Figure 5 illustrates the disparities in MSs perceived by 5 participants in 13 CUs when we simulate a multi-party Skype (Version 3.5.0.214) conversation (with HRD of 750 ms) using UDP traces of 5 sites (3 in N. America and 2 in Asia) collected in PlanetLab [12]. Because all traffic in Skype is routed through a common client, there are large disparities in MSs from one CU to the next.

Objective metrics of interactive VoIP. We define two objective metrics for characterizing the quality of multi-party VoIP [12]. These metrics can be tailored to two-party VoIP in a straightforward fashion.

a) The *conversational efficiency* (CE) measures the extension in time to accomplish a VoIP conversation when there are communication delays (Figure 2). Since a conversation over a network may be charged according to its duration, the same conversation will cost more for a network with lower CE.

$$CE = \frac{\text{Speaking Time} + \text{Listening Time}}{\text{Total Time of Call}}. \quad (1)$$

Note that CE is identically perceived by all participants.

TABLE I.
RESEARCH ISSUES IN INTERACTIVE VOIP SYSTEMS.

Sec. II: Evaluations of conversational quality
– Statistical model of subjective tests
– Trade-offs on objective metrics and JND
– Testbed for evaluating algorithms & systems
– Classifiers for learning and generalization
Sec. III: Cross-layer designs of speech codecs
– Codec design for dynamic encapsulation
– Rate-adaptive generation of enhancement layers
Sec. IV: Packet-stream-layer control algorithms
– Dynamic POS & LC control using classifiers
– Distributed equalization of MSs
Sec. V: Network-layer control algorithms
– Overlay networks and transport protocols
– Admissions control for multi-party VoIP

b) *Conversational symmetry* (CS). When a participant experiences highly asymmetric response times in a conversation, he/she tends to perceive a degradation in the naturalness of the conversation because it does not resemble a face-to-face conversation with uniform delays. One possible effect is, if A perceives B to be responding slowly, then A tends to respond slowly as well. To capture the asymmetry of MSs perceived by A, we define CS_A to be the ratio of the maximum and the minimum MSs experienced by A in a past window when the conversation switches from i to j ($i \rightarrow j$), and j is the speaker:

$$CS_{A(mp)} = \frac{\max_j MS_A^{i\rightarrow j}}{\min_{j,j\neq A} MS_A^{i\rightarrow j}} \text{ (multi-party VoIP)}$$

$$CS_{A(2p)} = \frac{\max_j MS_A^{i\rightarrow j}}{\min_j MS_A^{i\rightarrow j}} \text{ (2-party VoIP).} \quad (2)$$

CS_A is approximately 1 in a face-to-face conversation, but increases as the round-trip delay increases. In the two-party case, the minimum MS is always experienced by the speaker and the maximum experienced by the listener. Since there are multiple listeners in the multi-party case and the majority of the clients are passive listeners, it is important to identify the asymmetry for the passive listeners alone. Hence, we choose to eliminate the current speaker when evaluating the minimum MS in the multi-party case. Note that CS and CE are counteracting: as CS improves, CE degrades.

In VoIP, a user does not have an absolute perception of MEDs because he/she does not know when the other person will start talking or who will speak next (in multi-party VoIP). However, by perceiving the indirect effects of MED, such as MS and CE, the participants can deduce the existence of MED.

Figure I summarizes the four research issues in the design of interactive VoIP systems. These are related to the classification and generalization of subjective test results and the design of network-control and coding algorithms. For each of these issues, we present some existing work and our approaches in Sections II-V. Lastly, Section VI concludes the paper.

II. EVALUATING CONVERSATIONAL QUALITY

In this section, we first survey existing metrics for measuring conversational quality. We then present results on evaluating subjective quality of VoIP systems and methods for learning the mapping from objective metrics to control algorithms that optimize subjective quality.

A. Previous Work

Effects of MED on conversational quality. Subjective tests by Brady [11] and Richards [8] in the 1970s have led to the conclusions that MED affects the user perception of conversational quality, and that longer MEDs increase the dissatisfaction rate. However, their conclusions are limited when used for evaluating VoIP systems, since only a few constant delays were experimented. Subjective tests by Kiatawaki and Itoh [9] at NTT show that one-way delays are detectable, with a detectability threshold of 100-700 ms for trained crew and of 350-1100 ms for untrained subjects. ITU G.114 [13] prescribes that a *one-way delay* of less than 150 ms is desirable in voice communication, and that a delay of more than 400 ms is unacceptable. Without specifying the trade-offs with LOSQ, MED alone is not adequate for evaluating VoIP.

Objective measures on conversational quality. The International Telecommunication Union (ITU) has several recommendations for the objective and subjective evaluations of the end-to-end quality of a voice transmission system. Table II shows the naming standard established by ITU for the evaluation of the telephone transmission quality [14]. There are several recommendations for evaluating the objective conversational quality of a system in ACR (*absolute category rating*).

a) *PESQ* (ITU P.862) is an objective measure for evaluating speech quality based on the original and the degraded waveforms. It has been shown to have high correlations to subjective MOS results for a variety of land-line, mobile and VoIP applications. Because it only assesses the LOSQ but not the effects of delay, it must be used in conjunction with other metrics when evaluating conversational quality.

b) The *E-Model* (ITU G.107) was designed for estimating conversational quality in network planning. It considers the effects of the codec, packet losses, one-way delay, and echo. It is over-simplifying because it assumes the independence and additivity of degradations due to LOSQ and delay. Despite a number of extensions [15]–[19] that try to address its limitations, it is difficult to extend its role beyond network planning and use it for evaluating conversational quality in actual systems.

c) The *Call Clarity Index* (ITU P.561 and P.562) was developed for estimating the customer opinion of a voice communication system in a way similar to the E-Model. Although it provides models for PSTN systems, it does not have a user opinion model for packet switched networks with long delays and with non-linear and time variant signal processing devices, such as echo control and speech compression.

At this time, there is no single objective metric that can adequately capture the trade-offs among the factors that affect subjective conversational quality under all network and conversational conditions.

TABLE II.
ITU P.800.1 TERMINOLOGY ON TELEPHONE TRANSMISSION QUALITY.

Methodology	Listening-Only Conditions Tested	Conversational Conditions Tested
Subjective	MOS_{LQS}: P.800 Listening-only Tests	MOS_{CQS}: P.800 Conversational Tests
Objective	MOS_{LQO}: P.862 PESQ	MOS_{CQO}: P.562 for PSTN, not defined for VoIP
Estimated	MOS_{LQE}: Not defined	MOS_{CQE}: G.107 E-model

Subjective measures on conversational quality. A user's perception of a speech segment mainly depends on the intelligibility of the speech heard because the user lacks a reference to the original segment. To assess subjective conversational quality, formal mean-opinion-score (MOS) tests (ITU P.800) [14] are usually conducted. The method asks two subjects to complete a specific task over a communication system, ranks the quality using an ACR, and averages the opinion of multiple subjects.

There are several shortcomings of this approach for evaluating VoIP. Firstly, when completing a task and evaluating the quality of a conversation simultaneously, the cognitive attention required for both may interfere with each other. Secondly, the type and complexity of the task affects the quality perception. Tasks requiring faster turn taking can be more adversely affected by transmission delays than others. Thirdly, there is no reference in subjective evaluations, and ACR highly depends on the expertise of the subjects. Lastly, the results are hard to repeat, even for the same subjects and the same task.

In the NTT study [9] discussed earlier, subjective conversational experiments were conducted between two parties using a voice system with adjustable delays. Since the study did not consider the effect of losses and variations in delay, it is not applicable for VoIP systems.

ITU-T Study Group 12 has realized a lack of methods for evaluating conversational speech quality in networks and is currently conducting a study. However, it is not clear if the study will lead to an objective or a subjective methodology and whether the results can help design better VoIP systems.

Next, we describe our approaches to address the issues.

B. Evaluations/Generalization of Conversational Quality

Testbed for evaluating VoIP systems. We have developed a testbed for emulating two-party [6] and multi-party [12] VoIP. This entails the collection of Internet packet traces and multi-party interactive conversations and the design of a system to replay these traces and conversations. The prototype allows subjective tests to be repeated for different VoIP systems under identical network and conversational conditions [20].

The prototype consists of multiple computers, each running the VoIP client software, and a Linux router for emulating the real-time network traffic [6], [12]. We have modified the kernel of the router in order to intercept all UDP packets carrying encoded speech packets between any two clients. The router runs a troll program that drops or delays intercepted packets in each direction according to packet traces collected in the PlanetLab. We have also developed a human-response-simulator (HRS) that runs on each end-client. The HRSs simulate a conversation

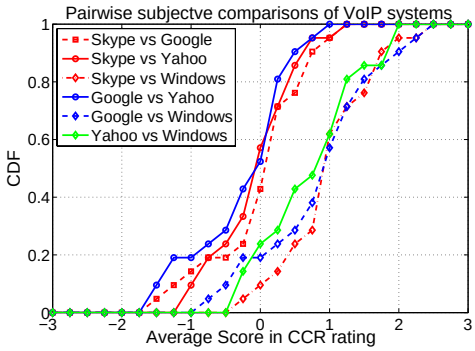

Figure 6. Distribution of pairwise subjective scores of four VoIP systems.

with pre-recorded speech segments by taking turns speaking their respective segments. We use a software interface to digitally transfer the waveforms to and from the clients without quality loss.

Subjective evaluations of four VoIP systems. We have compared four two-party VoIP clients: Skype (3.6), Google-Talk (beta), Windows Live Messenger (8.1), and Yahoo Messenger (8.1) [20]. Using conversations recorded by our testbed under some network and conversational conditions, human subjects were asked to comparatively evaluate two conversations by the CCR scale. The tests were conducted using six Internet traces under different network conditions and an additional trace representing an ideal condition with no loss and delay. We use three distinct conversations of different single-talk durations, HRD, and switching frequencies. The subjective test results in Figure 6 illustrate that Windows Live is preferred over the others. These are consistent with the objective metrics in terms of PESQ, CS, and CE (not shown). Similar tests have also been conducted to compare the multi-party version of Skype and our proposed system [12].

Statistical offline subjective tests. We have studied the statistical scheduling of offline subjective tests for evaluating alternative control schemes in real-time multimedia applications. These applications are characterized by multiple counteracting objective quality metrics (such as delay and signal quality) that can be affected by various control schemes. However, the trade-offs among these metrics with respect to the subjective preferences of users are not defined. As a result, it is difficult to select the proper control parameter value(s) that leads to the best subjective quality at run time. Since subjective tests are expensive to conduct and the number of possible control values and run-time conditions is prohibitively large, it is important that a minimum number of such tests be conducted offline, and that the results learned can be generalized to unseen conditions with statistical confidence. To this end, we have developed efficient

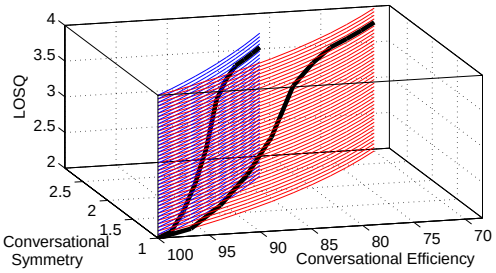

Figure 7. The points on each operating curve are a function of the MED imposed. Each plane represents some network and conversational conditions.

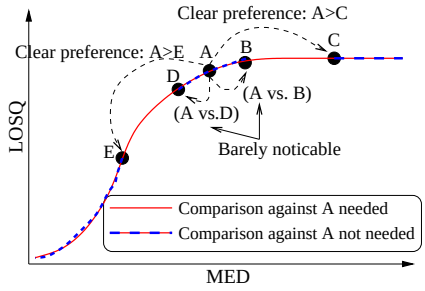

Figure 8. JND helps reduce the number of subjective evaluations

algorithms for scheduling a sequence of subjective tests under given conditions. Our goal is to minimize the number of subjective tests needed in order to determine the best point for operating the multimedia system to within some prescribed level of statistical confidence. A secondary goal is to efficiently schedule subjective tests under a multitude of operating conditions. Its success is based on the fact that humans can differentiate two such conversations when they are beyond the *just-noticeable difference* (JND, aka *difference limen*) [21]. Here JND is a difference in the physical sensory input that results in the detection of the change 50% of the time.

For a 2-party conversation, we use the following triplet in 3-D space to denote the operating point under a given codec and some network and conversational conditions:

$$CQ_{\text{2-party}} = \{LOSQ(MED, R), CE(MED), CS(MED)\}$$

where each axis represents an objective metric measured over a past window, and R is the redundancy degree. Figure 7 depicts the trade-offs as a function of MED and R for two conversations of different HRDs and switching frequencies [6]. For a conversation under some given conditions, the trade-offs between CE and CS are shown as a plane parallel to the LOSQ axis. Under these conditions, the possible LOSQs as a function of MED are shown by an *operating curve* on this plane.

The trade-offs shown by each operating curve are very complex and cannot be represented in closed forms because they involve some network and conversational conditions that cannot be modeled. Finding the *most probable* operating point on each curve with the best subjective quality (by selecting a proper MED) proves to be difficult because there are infinitely many operating points and each involves subjective tests. Also, the operating points do not have a total order because it may not always be possible to compare two conversations, one with high LOSQ but low CS and another with high CS but low LOSQ. To this end, we use JND as a vehicle to prune operating points with slightly different conditions that cannot be distinguished.

We have developed a method that uses the JND framework to discretize an operating curve in Figure 7 into a finite and manageable set [6]. Figure 8 illustrates the pruning of points on an operating curve that do not need to be compared against point A. Using subjective tests based on *comparison MOS*, which gives a relative subjective comparison of two conversations (similar to the Comparison Category Rating−CCR−method in ITU P.800 Annex E [14]), we have carried out simulated conversations and have conducted pair-wise subjective tests to capture the relative user preferences.

It is difficult to find the best operating point on an operating curve by comparing MOS scores because there are infinitely many such scores to be evaluated. Moreover, some operating points do not have to be assessed to the same confidence level when they are obviously inferior or unnecessary. To this end, we have developed a statistical method that minimizes the total number of tests for finding an operating point with the best subjective conversational quality to within a prescribed confidence level [22]. This is possible because statistical tests for comparing two conversations by human subjects follow a multinomial distribution.

When comparing two points on an operating curve, we have developed axioms on reflecivity, independence, identical statistical distribution, symmetry, indistinguishability, incomparability, and subjective preference. By using the axioms, we have constructed a general model for comparing two points on an operating curve that allows us to determine a likely direction on the location of the local optimum of subjective preference. However, the non-parametric nature of the model makes it difficult to combine the result of a test with the prior information obtained. Hence, we have also developed a parametric model of subjective comparisons after simplifying the general model. The simple model allows a probabilistic representation of our knowledge on the location of the local optimum and a way to statistically combine the deductions from multiple comparisons [22]. It also allows us to develop an adaptive search algorithm that significantly reduces the number of comparisons needed for identifying the local optimum. In addition, an estimate on the confidence of the result provides a consistent stopping condition for our algorithm.

Our results show that sequential evaluations of a single operating curve are the most effective in terms of minimizing the number of tests performed for that curve when identifying a local optimum to within some statistical confidence. Our simulation results show a substantial reduction in the number of comparisons by using a stopping criterion based on a lower confidence level, while incurring a negligible error in the estimation.

Based on this algorithm, the optimal strategy to min-

imize the total number of subjective tests for a set of operating curves is to test each curve sequentially and all the curves in parallel. In this approach, each subject is presented with a set of operating points to be compared, one from each operating curve to be tested. The tests in each set can be performed in any order and independent of other subjects because the result of the comparisons from one operating curve does not depend on that of another curve. At the end of the tests, the results from all the subjects are combined in order to generate a local optimum estimate and identify the next pair of operating points to be compared for each of the operating curves. As the number of operating curves to be tested is large, this approach allows subjects to independently carry out a batch of independent tests, without having to synchronize their results in a locked-step fashion with other subjects. The number of iterations is bounded by the typically small number of iterations to identify a local optimum candidate of an operating curve.

Classifiers for learning evaluation results. To address the issue that there are infinitely many possible network and conversational conditions, we propose to develop a classifier [23] that learns from training examples generated under limited conditions and that generalizes to unseen conditions.

Based on the pairwise comparisons of the conversations recorded on the four VoIP systems discussed last, we have generated training patterns, each consisting of 22 objective measures and a subjective measure. We have then learned these mappings by a classifier implemented as a support vector machine (SVM) [24] with a radial basis kernel function [20]. To simplify learning, we map the average of the user CCR opinions of A against B into 3 classes: A better than B, B better than A, and about the same. To verify that the classifier can generalize to unseen network and conversational conditions, we use cross validation techniques commonly employed in statistics. Our results show that we can predict 90% of the samples successfully in our training set and 70% of the cases when using 10-fold cross validation.

We have further used the subjective test results in the design of control algorithms that work well under a variety of conditions observed at run-time. The idea is to collect training data on subjective conversational quality offline and to design a classifier that learns the mappings between objective metrics measured in a past window and the control parameter value that leads to an operating point with the best subjective quality [25]. Based on a comprehensive set of network and conversational conditions, the training data is obtained by simulating two-way and multi-way VoIP conversations using our testbed. The simulations are carried out under each of the given conditions and values of the system-controllable parameters of the POS and LC algorithms, such as MED, redundancy degree, and level of MS equalization. Each element of the training set, therefore, consists of a mapping from the system-controlled parameter values and the objective metrics of the simulated conditions on a pair of conversations to their subjective preference. This method ensures that the conversations compared only differ by one parameter value and that their subjective preference can be attributed to the system-controlled value that leads to that opinion. We then learn a SVM classifier using training data based on the results of the subjective tests and the conditions under which the tests are conducted.

At run-time, the parameters representing the current conditions are estimated and input to the SVM. For example, in the design of the POS algorithm for two-party VoIP, loss, delay and jitter parameters are used to represent network conditions, and switching frequency and singe talk duration parameters represent conversational conditions. The SVM learned outputs the subjective preference for a given pair of points on the operating curve that corresponds to the network and conversational conditions observed. Its predictions on the subjective preference between multiple pairs of points on the same operating curve are combined using the statistical method described earlier in order to identify the optimal MED value, which is then used by the POS algorithm to adjust the jitter-buffer delay in order to achieve the operating point with the highest subjective quality.

III. Cross-Layer Speech Codecs for VoIP

Traditional codecs developed for cellular communications and PSTN calls are not suitable for VoIP because they have been designed for circuit switching under low bandwidth, fixed bit rates, and random bit errors. These codecs are not effective in packet-switched networks, whose loss rates and delay jitters are dynamic. Some recent codecs have been developed for VoIP applications. They can encode wide-band speech and exploit trade-offs between bit rate and delay in order to be more robust against bursty losses. However, they have been designed without due consideration of LC strategies in other layers of the protocol stack. Without such considerations, the LC strategies in these codecs can be inadequate and give sub-par performance, or redundant and unnecessary.

In this section, we first briefly survey speech codecs designed for VoIP. We then present the design of cross-layer speech codecs that are done in conjunction with LC strategies in the packet-stream layer.

A. Previous Work on Speech Codecs

Speech codecs were traditionally designed for applications in cellular and PSTN communications. With the proliferation of IP networks, they have been increasingly used in VoIP. They can be classified based on their coding techniques. *Waveform* codecs, such as ITU G.711 and G.726 [13], were designed to reconstruct a sample-wise waveform as closely as possible. *Parametric* codecs, such as G.722.2, G.723.1, G.728 and G.729A [13], model the production of speech in order to reconstruct a waveform that perceptually resembles the original speech. *Hybrid* codecs, such as G.729.1 [13], iLBC [44] and iSAC [45], combine techniques from both. Under no-loss conditions, the perceptual quality of a codec is a function of its coding

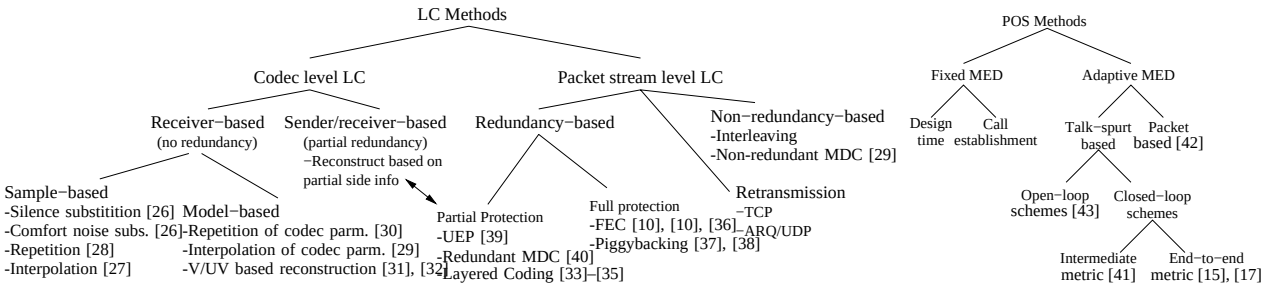

Figure 9. A classification of existing LC and POS methods at (a) the codec and (b) the packet-stream layers.

technique and bit rate. However, it is difficult to compare codecs under loss conditions.

Parametric and hybrid codecs are popular in VoIP because they have lower bit rates and better perceptual quality. By controlling the frame size and the frame period, their design involves trade-offs among robustness, quality, and algorithmic delay [46]. Frame size generally varies between 10 bytes (G.729A) and 80 bytes (G.729.1, 32-kbps wideband mode), with multi-mode codecs having a wide range. (For example, G.722.2 has frames between 17 and 60 bytes.) Frame period varies between 10 ms (G.729A) and 60 ms (iSAC with 30-60 ms adaptive size). The most common periods are 20 ms (such as iLBC 15.2-kbps mode, G.729.1 and G722.2) and 30ms (iLBC 13.3-kbps mode and G723.1). In general, a larger frame with a shorter period achieves higher quality within the multiple modes of a codec or within a family of codecs with similar coding techniques. However, a longer frame and look-ahead window may incur more algorithmic delay.

Figure 9a summarizes some of the LC techniques employed in speech codecs. Receiver-based schemes that do not use redundant information can be classified into sample-based and model-based. In early systems, silence or comfort-noise substitution [26] or repetition of the previous frame [28] was proposed in place of a lost frame. Also proposed was the transmission of even and odd samples in separate packets and using sample-based interpolation when a packet was lost [27]. Early model-based schemes simply repeat the codec parameters of the last successfully received frame [30]. Later, interpolation of codec parameters from the previous and the next frame was proposed [29]. Other schemes utilized the information about the voiced/unvoiced properties of a speech frame to apply specialized LC for reducing the perception of degradation [31], [32]. Schemes that require the cooperation of the sender and the receiver utilize partial redundant information [33]–[35] that is made available by the packet-stream-level LC (see Section IV).

The trade-offs between frame size and frame period are different from those between packet size and packet period in the packet-stream layer. To avoid excessive losses in the Internet, it is important to choose an appropriate packet period, as long as packets are smaller than an MTU of 576 bytes and can be sent without fragmentation [47]. (In practice, an MTU of 1500 bytes will not cause fragmentation.) For VoIP using IPv4, a packet period between 30 ms and 60 ms generally works well. When the frame period is shorter than the packet period, multiple frames have to be encapsulated in a packet before they are sent. For some codecs, the loss of a single packet can cause a misalignment of its internal states and degrade its decoded output. For example, a single lost frame in G.729A [48] can lead to perceptible distortions in the reconstructed speech for up to 20 frames [33], [34].

Although the large MTU in packets relative to the frame size provides new opportunities for LC in the packet-stream layer (see Section IV), the LC mechanisms in the packet-stream and the codec layers may not work well together. This happens because techniques for recovering lost frames [49] in codecs often exploit information in the coded speech [50] that may not be properly encapsulated into packets. For instance, G.729.1 [35], the wide-band version of G.729A, recovers a single lost frame using multiple layers of information from the previous and the next frames received. This LC technique will be not useful when multiple adjacent frames are encapsulated into a single packet that is lost. For this LC technique to be effective, the information for reconstructing a lost frame must be encapsulated in different packets.

The Internet Low Bit-rate Codec (iLBC) [44] used in early versions of Skype and its extensions [51], [52] address this issue by encoding frames into self-decodable units using a modified CELP coder with increased bit rate. Although this approach avoids the propagation of internal-state errors after a loss, distortions are still perceptible unless additional LC mechanisms are implemented in the packet-stream layer.

Recently, Global IP Sound (GIPS) has released the second version of its proprietary iSAC wideband speech codec for use in Skype. Its white paper [45] indicates that iSAC uses an adaptive packet period of 30-60 ms, with an adaptive bit rate of 10-32 kbps and a separate low complexity mode. Although the white paper claims that the codec achieves better performance than G.722.2 for comparable bit rates, there is no independent validation of the claim and no information on its LC capability.

In evaluating a speech codec designed, its performance is commonly evaluated by comparing the quality of its reconstructed waveforms under ideal and non-ideal network conditions. One common method of generating non-ideal network conditions is to use stochastic models.

Sun and Ifeachor proposed a simple Bernoulli model with independent packet losses for modeling the loss behavior in VoIP [15]. The model is highly approximate

because Internet packet losses exhibit temporal dependencies [53], especially for periodic transmissions. Further, speech quality can vary significantly across different loss patterns with the same average rate [10]. A second approach based on the Gilbert model has been used for modeling the loss behavior of Internet traces in IP telephony [10] and multimedia [53] and for evaluating speech codecs [51], [52]. The model is approximate because it assumes that a packet loss only depends on the loss of the previous packet. Extended models, such as the n-state Markov chain and the extended Gilbert model [53], use additional parameters to model the dependency of losses.

The main deficiency of these models is that they do not consider the LC algorithm in the packet-stream layer, such as redundant piggybacking [34] and multi-description coding (MDC) [49], [54], [55]. There are a number of recent studies on cross-layer designs of codecs [56]–[60], but none has focused on designs for real-time VoIP. To provide efficient use of resources and the best conversational quality, the LC mechanisms in the packet-stream and the codec layers must be developed in a coupled fashion.

To optimize perceptual quality, our survey identifies the need to design the LC mechanism in codecs with that of the packet-stream layer. This means that the encoding of speech into frames must dynamically adapt to the packet rate, which in turn adapts to network congestion. In the next section, we describe our approach to improve the design of codecs for VoIP.

B. Cross-Layer Designs of Speech Codecs

Codecs with self-decodable units. To avoid the propagation of errors in internal states across packet boundaries and to maximize coding efficiency within a packet, we have designed codecs that encode frames in such a way that are self-decodable but may have dependent internal states when encapsulated into a packet. This is similar to what was done in iLBC [44]. In addition, these codecs can operate in multiple modes, in terms of frame size and packet period selected by the SVM learned in the packet-stream layer for optimizing conversational quality.

Based on the ITU G.729 CS-ACELP speech codec operating at 8 kbps, we have developed cross-layer designs with redundancy-based LC. In our first design [33], we have increased the frame length in order to reserve space for redundancies at the packet level, without increasing the bit rate. It uses MDC to conceal losses at the packet level and adapts to dynamic loss conditions, while maintaining an acceptable end-to-end delay. By protecting only the most important excitation parameters of each frame according to its speech type, our approach enables more efficient usage of the bit budget. In our second design [34], we have developed a variable bit-rate layered coding scheme that dynamically adapts to the characteristics of the speech encoded and the network loss conditions. To cope with bursty losses while maintaining an acceptable end-to-end delay, our scheme employs layered coding with redundant piggybacking of perceptually important parameters in the base layer, with a degree of redundancy adjusted according to feedbacks from receivers. Under various delay constraints, we study trade-offs between the additional bit rate required for redundant piggybacking and the protection of perceptually important parameters. Although these cross-layer designs incorporate LC information in the packet-stream layer, G.729 is not the perfect codec because it suffers from the propagation of errors in internal states across packet boundaries. We have also applied a similar approach to the design of the G.722.2 and G.729.1 wide-band speech codecs that have self-decodable frames.

Cross-layer designs of speech codecs. To facilitate the generation of information for effective LC, the encoder needs to know the amount of payload in each packet available for carrying the LC information. This is important in multi-party VoIP when multiple voice streams have to be encapsulated in the same packet and the payload for carrying redundant information is limited. In this case, the codec needs to generate enhancement layers, and the piggybacking of enhancement layers depends on the payload available.

Codec design for dynamic encapsulation. We have developed a codec that allows the flexible encapsulation of frames into packets with self contained internal states and that avoids the propagation of internal-state errors across packet boundaries. We have modified the G.722.2 codec that can operate in a wide range of frame periods (10-60 ms) and bit-rates (8-40 kbps). Based on the information provided by the packet-stream layer of the VoIP client, the coder can switch among different modes during a conversation in order to adapt to changes in the network and conversational conditions. The packet period used is propagated to the coder to ensure that it removes any redundancies among the multiple frames encapsulated in a packet. A similar adaptation has been observed in Skype, although its mechanism cannot be conclusively deduced to changes in network conditions, as its source code is unavailable.

Off-line learning of codec's robustness to LC. We have extend the learning of classifiers in the packet-stream layer to accurately predict LOSQ at run time, using unconcealed frame patterns available at the receiver [6]. The advantage of this approach over PESQ is that it does not require the original speech frames in the computation (which are needed for the calculation of PESQ). We have evaluated the robustness of our codec against a variety of loss conditions using off-line experiments. Network traces collected were processed using several different LC and POS schemes, and their unconcealed frame patterns were generated. Speech segments were then be encoded and decoded using various unconcealed frame patterns, and the output speech was objectively evaluated by PESQ. Finally, a classifier was trained using the unconcealed frame patterns and the corresponding PESQs.

IV. Loss Concealments & Play-out Scheduling

The packet-stream layer mitigates delays, losses, and jitters at the packet level through its LC and POS algorithms. As is discussed in the last section, these algorithms should be designed in conjunction with the speech codec and with an understanding of conversational quality. In this section, we identify the limitations of existing LC and POS algorithms. We then discuss our approaches to address the issues in their design.

A. Previous Work

Figure 9 summarizes some existing LC techniques at the packet-stream layer. They aim to either reduce the amount of unconcealable frames experienced by the decoder or provide partial redundancy for helping the decoder reduce perceptual degradations due to losses.

Retransmissions of speech frames after the detection of a network loss is infeasible in real-time VoIP, due to the excessive delays involved and their effects on MED.

Non-redundant LC schemes are generally based on the interleaving of frames during packetization [61]. One way is to exploit the fact that shorter distortions are less likely to be perceived, and to break an otherwise long segment into several shorter segments that are close by, but not consecutive. This is not strictly an LC technique because it does not actually recover losses. Another way is MDC [49], [54], [55] that generates multiple descriptions with correlated information from the original speech data. This may be hard in low bit-rate streams whose correlated information has been largely removed during coding [49]. Another disadvantage is that the receiver will incur a longer MED when waiting for all the descriptions to arrive before declaring a description is lost.

Redundant LC schemes exploit trade-offs among the redundancy level, the delay for recovering losses from the redundant information, and the quality of the reconstructed speech. They work in the Internet because increases in packet size, as long as they are less than the MTU [47], do not lead to noticeable increases in the loss rate [38]. They consist of schemes that use partial and full redundancies. Examples employing partial redundancies include layered coding [33]–[35], UEP (unequal error protection) [39], and redundant MDC [40]. Examples employing full redundancies include FEC (forward error correction) [10], [10], [36] and redundant piggybacking [37], [38]. An FEC-based LC scheme [17] for VoIP incorporates into its optimization metric the additional delay incurred due to redundancy. In our previous work, we have used piggybacking as a simple yet effective technique for sending copies of previously sent frames together with new frames in the same packet, without increasing the packet rate [6], [12], [38].

The main difficulty of using redundant LC schemes is that it is hard to know a suitable redundancy level. Its dynamic adaption to network conditions may either be too slow, as in Skype [38], or too conservative [6]. Another consideration is that the redundancy level is application-dependent. Fully redundant piggybacking is suitable in two-party VoIP, but partial redundancy may need to be used in multi-party VoIP when speech frames from multiple clients are encapsulated in the same packet.

Figure 9b also summarizes the various POS methods. Due to non-stationary and path-dependent delays and losses, simple schemes with fixed MEDs either hardcoded at design time or during call establishment do not provide consistent protection against late losses. Adaptive POS schemes that adjust the playout schedule at the talk-spurt or the packet level are more prevalent.

At the talk-spurt level, silence segments can be added or omitted at the beginning of a talk spurt in order to make the changes virtually imperceptible to the listener. Adjustments can also be made for each frame using *time-scale modification* [42] that stretches or compresses frames without changing its pitch period. However, it requires additional computational resources, has small effects on MEDs, and is generally perceptible.

At the packet level, there have been several studies that aim to balance between the number of packets late for playout and the jitter-buffer delay that packets wait before their scheduled playout times. *Open-loop schemes* use heuristics for picking some system-controllable metrics (such as MED), based on network statistics available [43]. They are less robust because they do not explicitly optimize a target objective. Moreover, they do not consider the effects of the codec on speech quality, although their performance depends on the codec used. *Closed-loop schemes with intermediate quality metrics* [41] control an intermediate metric based on the late-loss rate collected in a window. Their difficulty lies in choosing a good intermediate metric. *Closed-loop schemes with end-to-end quality metrics* generally use the E-model [1] for estimating conversational quality as a function of some objective metrics. One study uses this estimate in a closed-loop framework to jointly optimize the POS and FEC-based LC [17]. Another study [15] proposes to use the E-model but separately trains a regression model for modeling the effects of the loss rate and the codec on PESQ. These models are limited because, without a redundancy-based LC scheme, lost frames cannot be recovered by adjusting the playout delays alone.

Existing VoIP systems usually employ redundancy-based LC algorithms for recovering losses when using UDP. However, none of these approaches considers delay-quality trade-offs for delivering VoIP of high perceptual quality to users. Previous LC algorithms based on analytic loss models [41], [43] do not always perform well, as these models may not fully capture the dynamic network behavior and do not take into account the LC strategies in codecs. Existing POS algorithms based on open-loop heuristic functions [43] may not be robust under all conditions, whereas closed-loop approaches [41] are difficult to optimize without a good intermediate metric. Some recent approaches [15], [17] have employed an end-to-end objective metric, such as the E-model, as their intermediate metric. There is also very little reported results on POS algorithms for multi-party VoIP [62].

We present in the next section new LC and POS control algorithms that address the trade-offs related to conversational quality. Using the classifier learned, we use run-time network and conversational conditions to select the best operating point of these control algorithms. A related problem studied is the equalization of MSs for improving perceptual quality in multi-party VoIP. We also consider the design of these algorithms with the design of the codec and the network-control algorithms in multi-party VoIP.

B. Packet-Stream Layer LC and POS Algorithms

Two-party LC and POS schemes. We have developed new POS/LC schemes for dynamically selecting a playout schedule for each talk-spurt [6] and an appropriate redundancy degree for each packet. Using the loss information of the 100 most recent packets (≈3 sec) from the receiver, the sender selects a redundant piggybacking degree in order to achieve 2% target unconcealed packet loss rate. Our POS scheme at the receiver uses delay information, redundant piggybacking degree, and predictions of objective metrics (LOSQ, CS and CE) for the upcoming talk-spurt in order to select a suitable playout schedule for that talk-spurt. Since the conversational condition changes slowly during a conversation, CS and CE can be accurately estimated by monitoring the silence and the voiced durations at the receiver. In estimating the LOSQ curve, we first conduct offline experiments to learn a classifier that maps network conditions to the corresponding PESQs. By considering bursty loss patterns in real traces, our approach leads to significantly more accurate estimates of PESQ when compared to those in the previous work [15] with IID loss patterns. Our classifier is then used at run time to estimate the relation between MED and LOSQ.

Figure 10 depicts the decisions made by our control schemes for two connections with large jitters and medium losses. It shows that our schemes can closely track the changing network conditions, while making discrete adjustments when needed in order to keep the conversational quality in a user preferred state. Our system also performs better than the p-optimum algorithm [15], which estimates conversational quality by a hybrid E-Model and PESQ [6].

Lastly, we have integrated the classifiers learned into the design of a prototype VoIP system. The classifiers enable the systematic tuning of the control algorithms in our prototype, without the need to carry out expensive subjective tests.

Multi-party LC and POS schemes. We observe that, from a client's (say A) perspective, the decisions made by the POS of other clients in equalizing MSs in the current turn do not affect the MS observed by A. Hence, A's decision can be assumed to be independent of the concurrent decisions made by the other clients. Figure 11 extends the trade-off curve for the two-party VoIP in Figure 7 to the multi-party case. It depicts two trade-off

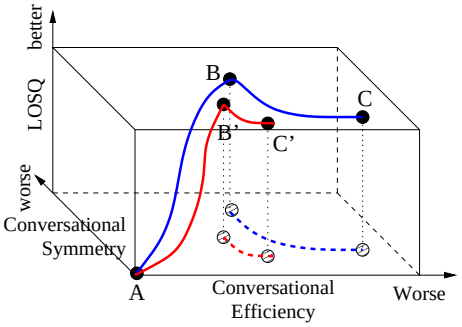

Figure 11. Trade-offs under different multi-party conditions.

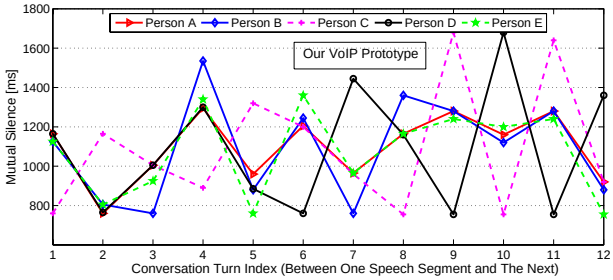

Figure 12. MSs experienced in our VoIP prototype.

curves: the curve connecting A, B, and C corresponds to a network condition with high disparities in delays among the connections; and the curve connecting A, B' and C' corresponds to a condition with similar average delay but considerably less disparities. The control from A to B (*resp.*, A to B') is similar to the two-party case: increasing the MED towards B (*resp.*, B') conceals more packets and improves LOSQ but degrades CS and CE. In the multi-party case, further increasing the MED from B to C (*resp.*, B' to C') to achieve full equalization will lead to a high LOSQ with improved CS but degraded CE. Hence, operating at C will result in a highly inefficient conversation with low conversational efficiency; whereas operating at C' is relatively more efficient than at C.

We have developed control schemes that operate in conjunction with an overlay network (described in the next section) in multi-party VoIP [12]. To reduce the overhead and to improve loss adaptations, our LC scheme operates on a link-by-link basis to select the redundant piggybacking degree. When multiple speakers are talking, we combine their packets destined to the same client into one (without exceeding the MTU). Although our scheme requires each node to maintain retransmission buffers for storing recently received frames, the overhead is manageable when the piggybacking degree is 4 or less.

Our POS scheme is different than the two-party counterpart. In the multi-party case, the order of the speakers is unknown, and the network conditions among the participants may have large disparities. We address this problem by equalizing the MSs observed by different listeners in the same turn. For listeners whose MED affects the efficiency of the whole conversation (bottleneck clients), we use stricter MED values that closely hugs the delay curve, similar to that in Figure 10a. For other listeners, we use less strict MEDs. In comparison to the MSs in Skype in Figure 5, Figure 12 shows that our system has

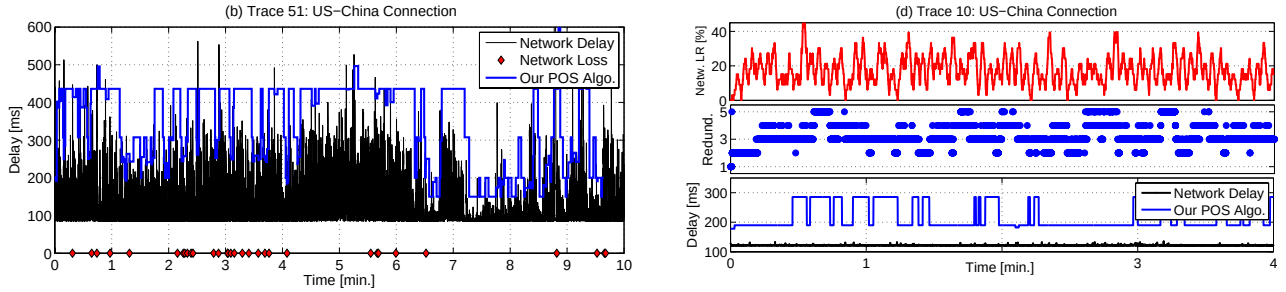

Figure 10. Network delays and POS/LC control decisions made by our JND-based POS and LC schemes for 2 connections: (a) connection with high and time-varying delay spikes; and (b) connection with medium packet losses.

lower average MS as well as less variations.

We are in the process of developing a general equalization algorithm with parameters that operate in continuous equalization levels [12]. Using simulated conversations generated under various control parameters and network and conversational conditions, we plan to conduct subjective JND tests to determine the perceptual sensitivity of humans in conversational quality as a function of the level of equalization. We will develop a statistical procedure similar to that described in Section II-B and experiment on some limited cases to within a prescribed confidence level. Finally, we plan to train a classifier offline and apply it at run time to determine the control parameters that will lead to the most preferred conversational quality. We expect our approach to be general and lead to better equalization than that in Figure 12.

V. Network-Layer Support for VoIP

In this section, we present related work on network-control algorithms for supporting VoIP.

Several standardized and proposed transport protocols, such as TCP, UDP and RTP, were designed with different trade-offs between reliable delivery of packets and application-layer end-to-end delay. By providing hooks for synchronization, reliability, QoS feedback, and flow control, RTP can support real-time multimedia applications [63]–[65]. Although RTP does not guarantee the real-time delivery of continuous media, it relies on resource reservations protocols, such as RSVP [66], for resource scheduling. RTP can be used in conjunction with the mechanisms described in this paper to ensure high conversational quality. Another approach is to design end-to-end protocols that are more TCP friendly [67], [68]. These protocols will need to be extended in order to address the trade-offs in conversational quality.

In multi-party VoIP, the current speaker(s) needs to convey his/her speech information to all the listeners. A good design should accommodate dynamic and diverse network conditions among the clients. The protocol and the connection topology used are generally dictated by where audio mixing is done [69]. When a VoIP client (as in Skype [70]) or a bridge (as in QQTalk [71]) is responsible for decoding, mixing, and encoding the signals from the clients, it is natural for all the clients to send their packets to the centralized site to be mixed and forwarded. The approach may not be scalable because it can create a bottleneck near that site. Moreover, the *maximum end-to-end delay* (ME2ED) between any speaker-listener pair can be large when the clients are geographically distributed [12].

On the other hand, a distributed approach asks each client to independently manage its transmissions. One way is for each speaker to multicast the speech information to all listeners. Although multicasts are available in the Internet [72], the support of reliable real-time multicasts for receivers of different loss and delay behavior is very preliminary [73]. The focus in the IETF working group on a NACK-based asynchronous-layered coding protocol with FEC [74] is inadequate for multi-party VoIP.

A hybrid approach is to have an overlay network [75] that uses a subset of the clients to manage the mixing and forwarding of unicast packets. The approach achieves a shorter ME2ED than a centralized approach and a smaller number of unicast messages than a fully distributed approach. One issue, however, is that it is complex for the overlay clients to coordinate the decoding, mixing, and encoding of the speech signals. Alternatively, we have taken the approach for the overlay clients to simply encapsulate the speech frames from the multiple clients into a single packet before forwarding them [12]. This is not an issue as long as the number of streams to be encapsulated fits within the MTU of each packet.

To design a topology with proper trade-offs between ME2ED among the clients and the maximum number of packets relayed in a single packet period by any client, we have studied a commonly used overlay topology constructed from a subset of the clients (called parent nodes) [12]. The topology assumes that all the parent nodes are fully connected, and that each remaining node (called child node) is connected to only one parent node. When a call is set up, the client that initiated the call collects delay and loss information among the clients. Due to the prohibitive nature of enumerating all topologies, we use a greedy algorithm to find a good topology. The heuristic first determines the client pair with ME2ED (called bottleneck pair) in a fully connected topology. It then finds a single-parent topology that minimizes ME2ED. If the improvement in ME2ED is small (say less than 50 ms), then it uses the best single-parent topology as the overlay network; otherwise, it adds a second parent node. It iteratively increases the number of parents until either the difference between the ME2EDs of the current topology and the fully connected topology is small or the bottleneck pair in the current topology is directly

connected. The process can be repeated whenever there is a significant change in the network conditions.

We are in the process of developing admissions-control algorithms for deciding whether a new client can be added without adversely affecting the conversational quality of all the listeners, and the dynamic dissemination of network information for operating the control algorithms. We plan to initially use UDP as the transport protocol but will consider RTP and other protocols at a later stage.

VI. Conclusions

In this paper, we have presented some solutions on the study of real-time two-party and multi-party VoIP systems that can achieve high perceptual conversational quality. Our solutions focus on the fundamental understanding of conversational quality and its trade-offs among the design of speech codecs and strategies for network control, playout scheduling, and loss concealments.

The degradation in the perceptual quality of an interactive conversation over a network connection is caused by a combination of the decrease in speech quality when packets are lost or delayed, and the asymmetry in silence periods when the conversation switches from one speaker to another. Its study is largely unexplored because there is no single objective metric for assessing the quality of a VoIP conversation whose results match well with subjective results. Indiscriminate subjective testing is not feasible because it is prohibitively expensive to carry out many such tests under various conversational and network conditions. Moreover, there is no systematic method to generalize the subjective test results to unseen conditions. To this end, there are three issues studied in this paper that address the limitations of existing work and that improve the perceptual quality of VoIP systems.

- We have presented a statistical approach based on just-noticeable difference to significantly reduce the large number of subjective tests, as well as a classification method to automatically learn and generalize the results to unseen conditions. Using network and conversational conditions measured at run time, the classifier learned helps adjust the control algorithms in achieving high perceptual conversational quality.
- We have described the concept of a cross-layer speech codec to interface with the loss-concealment and playout scheduling algorithms in the packet-stream layer in order to be more robust and effective against packet losses.
- We have presented a distributed algorithm for equalizing mutual silences and an overlay network for multi-party VoIP systems. The approach leads to multi-party conversations with high listening-only speech quality and balanced mutual silences.

References

[1] ITU-G.107, "The E-model, a computational model for use in transmission planning," http://www.itu.int/rec/T-REC-G.107/en. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/i/ITUG107.pdf

[2] ITU-P.561, "In-service non-intrusive measurement device: Voice service measurements," http://www.itu.int/rec/T-REC-P.561/en. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/i/ITUP561.pdf

[3] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, pp. 696–735, 1974. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/h/HSEASGJ74.pdf

[4] K. Weilhammer and S. Rabold, "Durational aspects in turn taking," in *Proc. Int'l Conf. on Phonetic Sciences*, 2003. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/k/KWSR03.pdf

[5] L. T. Bosch, N. Oostdijk, and J. P. de Ruiter, "Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues," *Text, Speech, and Dialogue*, pp. 563–570, 2004.

[6] B. Sat and B. W. Wah, "Playout scheduling and loss-concealments in VoIP for optimizing conversational voice communication quality," in *Proc. ACM Multimedia*, Augsburg, Germany, Sept. 2007, pp. 137–146.

[7] ——, "Evaluation of conversational voice quality of the Skype, Google-Talk, Windows Live, and Yahoo Messenger VoIP systems," in *IEEE Int'l Workshop on Multimedia Signal Processing*, Oct. 2007.

[8] D. L. Richards, *Telecommunication by Speech*. Butterworths, UK-London, 1973.

[9] N. Kiatawaki and K. Itoh, "Pure delay effect on speech quality in telecommunications," *IEEE Journal on Selected Areas of Communication*, vol. 9, no. 4, pp. 586–593, May 1991. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/n/NKKI91.pdf

[10] J.-C. Bolot, S. Fosse-Parisis, and D. Towsley, "Adaptive FEC-based error control for Internet telephony," in *Proc. IEEE INFOCOM*, vol. 3, 1999, pp. 1453–1460. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/j/jcbsfpdt99.pdf

[11] P. T. Brady, "Effects of transmission delay on conversational behaviour on echo-free telephone circuits," *Bell System Technical Journal*, vol. 50, no. 1, pp. 115–134, Jan. 1971.

[12] B. Sat, Z. X. Huang, and B. W. Wah, "The design of a multi-party VoIP conferencing system over the Internet," in *Proc. IEEE Int'l Symposium on Multimedia*, Taichung, Taiwan, Dec. 2007, pp. 3–10.

[13] International Telecommunication Union, "ITU-T G-Series recommendations," http://www.itu.int/rec/T-REC-G/en.

[14] ——, "ITU-T P-Series recommendations," http://www.itu.int/rec/T-REC-P/en.

[15] L. Sun and E. Ifeachor, "New models for perceived voice quality prediction and their applications in playout buffer optimization for VoIP networks," in *Proc. IEEE Communications*, vol. 3, 2004, pp. 1478–1483.

[16] ——, "Voice quality prediction models and their applications in VoIP networks," *IEEE Trans. on Multimedia*, vol. 9, no. 4, pp. 809–820, 2006. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/l/LSEI06.pdf

[17] C. Boutremans and J.-Y. L. Boudec, "Adaptive joint playout buffer and FEC adjustment for Internet telephony," in *Proc. IEEE INFOCOM*, vol. 1, 2003, pp. 652–662. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/c/cbjylb03.pdf

[18] A. Takahashi, H. Yoshino, and N. Kitawaki, "Perceptual QoS assessment technologies for VoIP," *IEEE Communications Magazine*, pp. 28–34, July 2004. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/a/ATHYNK04.pdf

[19] A. P. Markopoulou, F. A. Tobagi, and M. J. Karam, "Assessing the quality of voice communications over Internet backbones," *IEEE/ACM Trans. on Networking*,

vol. 11, no. 5, pp. 747–760, 2003. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/a/APMFATMJK03.pdf

[20] B. Sat and B. W. Wah, "Evaluating the conversational voice quality of the Skype, Google-Talk, Windows Live, and Yahoo Messenger VoIP systems," *IEEE Multimedia*, (accepted) Dec. 2008.

[21] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. Springer, 1972.

[22] B. Sat and B. W. Wah, "Statistical testing of off-line comparative subjective evaluations for optimizing perceptual conversational quality in voip," in *Proc. IEEE Int'l Symposium on Multimedia*, Dec. 2008, p. (accepted to appear).

[23] P. L. Lanzi, "Learning classifier systems: Then and now," *Evolutionary Intelligence*, vol. 1, pp. 63–82, 2008.

[24] C.-C. Chang and C.-J. Lin, "A library for support vector machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm/. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[25] Z. X. Huang, B. Sat, and B. W. Wah, "Automated learning of play-out scheduling algorithms for improving the perceptual conversational quality in multi-party VoIP," in *Proc. IEEE Int'l Conf. on Multimedia and Expo*, July 2008, pp. 493–496.

[26] J. Suzuki and M. Taka, "Missing packet recovery techniques for low-bit-rate coded speech," *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 5, pp. 707–717, June 1989.

[27] N. S. Jayant and S. W. Christensen, "Effects of packet losses in waveform coded speech and improvements due to odd-even sample-interpolation procedure," *IEEE Trans. on Communications*, vol. 29, no. 2, pp. 101–110, Feb. 1981.

[28] R. C. F. Tucker and J. E. Flood, "Optimizing the performance of packet-switch speech," in *IEEE Conf. on Digital Processing of Signals in Communications*, Loughborough University, Apr. 1985, pp. 227–234.

[29] D. Lin, *Loss Concealments for Low Bit-Rate Packet Voice*. Urbana, IL: Ph.D. Thesis, Dept. of Electrical and Computer Engineering, Univ. of Illinois, Aug. 2002.

[30] S. Atungsiri, A. Kondoz, and B. Evans, "Error control for low-bit-rate speech communication systems," *IEE Proc. I: Communications, Speech and Vision*, vol. 140, no. 2, pp. 97–103, Apr. 1993.

[31] J. Wang and J. D. Gibson, "Parameter interpolation to enhance the frame erasure robustness of CELP coders in packet networks," in *IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, 2001, pp. 745–748.

[32] J. F. Wang, J. C. Wang, J. Yang, and J. J. Wang, "A voicing-driven packet loss recovery algorithm for analysis-by-synthesis predictive speech coders over Internet," *IEEE Trans. on Multimedia*, vol. 3, no. 1, pp. 98–107, Mar. 2001.

[33] B. Sat and B. W. Wah, "Speech-adaptive layered G.729 coder for loss concealments of real-time voice over IP," in *Proc. IEEE Int'l Conf. on Multimedia and Expo*, July 2005.

[34] ——, "Speech- and network-adaptive layered G.729 coder for loss concealments of real-time voice over IP," in *IEEE Int'l Workshop on Multimedia Signal Processing*, Oct. 2005.

[35] ITU-G.729.1, "G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729," http://www.itu.int/rec/T-REC-G.729.1/en. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/i/ITUG729.1.pdf

[36] N. Shacham and P. McKenney, "Packet recovery in high-speed networks using coding and buffer management," in *Proc. of IEEE INFOCOM*, May 1990, pp. 124–131.

[37] T. J. Kostas, M. S. Borella, I. Sidhu, G. M. Schuster, J. Grabiec, and J. Mahler, "Real-time voice over packet-switched networks," *IEEE Network*, vol. 12, no. 1, pp. 18–27, January-February 1998.

[38] B. Sat and B. W. Wah, "Analysis and evaluation of the Skype and Google-Talk VoIP systems," in *Proc. IEEE Int'l Conf. on Multimedia and Expo*, July 2006.

[39] M. Chen and M. N. Murthi, "Optimized unequal error protection for voice over IP," in *ICASSP*, 2004, pp. 865–868. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/m/MCMNM04.pdf

[40] W. Jiang and A. Ortega, "Multiple description coding via polyphase transform and selective quantization," in *Proc. Visual Communications and Image Processing*, vol. 3653, Dec. 1998, pp. 998–1008. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/j/JiaOrt98.ps.gz

[41] S. B. Moon, J. Kurose, and D. Towsley, "Packet audio playout delay adjustment: performance bounds and algorithms," *Multimedia Systems*, vol. 6, no. 1, pp. 17–28, Jan. 1998.

[42] Y. J. Liang, N. Faber, and B. Girod, "Adaptive playout scheduling and loss concealment for voice communication over IP networks," *IEEE Trans. on Multimedia*, vol. 5, no. 4, pp. 532–543, Dec. 2003.

[43] R. Ramjee, J. Kurose, D. Towsley, and H. Schulzrinne, "Adaptive playout mechanisms for packetized audio applications in wide-area networks," in *Proc. 13th Annual Joint Conf. IEEE Computer and Communications Societies on Networking for Global Commmunication*, vol. 2, 1994, pp. 680–688.

[44] S. Andersen, A. Duric, H. Astrom, R. Hagen, W. Kleijn, and J. Linden, "Internet low bit rate codec (iLBC)," Dec. 2004, http://www.ietf.org/rfc/rfc3951.txt. [Online]. Available: http://www.ietf.org/rfc/rfc3951.txt

[45] Global IP Sound, "Datasheet of GIPS Internet Speech Audio Codec," 2007, http://www.gipscorp.com/files/english/datasheets/iSAC.pdf. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/i/iSAC.pdf

[46] W. C. Chu, *Speech Coding Algorithms*. Wiley, 2003.

[47] IETF, "RFC 791, Internet Protocol: DARPA Internet program protocol specification," Sept. 1981, http://www.ietf.org/rfc/rfc791.txt.

[48] ITU-G.729, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)," http://www.itu.int/rec/T-REC-G.729/en. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/i/ITUG729.pdf

[49] D. Lin and B. W. Wah, "LSP-based multiple-description coding for real-time low bit-rate voice over IP," *IEEE Trans. on Multimedia*, vol. 7, no. 1, pp. 167–178, Feb. 2005.

[50] T. Chua and D. C. Pheanis, "QoS evaluation of sender-based loss-recovery techniques for VoIP," *IEEE Network*, pp. 14–22, Dec. 2006. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/t/TCDCP06.pdf

[51] C. M. Garrido, M. N. Murthi, and S. V. Andersen, "Towards iLBC speech coding at lower rates through a new formulation of the start state search," in *Proc. ICASSP*, vol. 1, 2005, pp. 769–772. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/c/CMGMNMSVA05.pdf

[52] ——, "On variable rate frame independent predictive speech coding: Re-engineering iLBC," in *Proc. ICASSP*, vol. 1, 2006, pp. 717–720. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/c/CMGMNMSVA06.pdf

[53] W. Jiang and H. Schulzrinne, "Modelling of packet loss and delay and their effect on real-time multimedia service quality," in *Proc. Int'l Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)*, 2000.

[54] A. A. E. Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. on Information Theory*, vol. 28, pp. 851–857, Nov. 1982.

[55] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Processing Magazine*, pp. 74–93, Sept. 2001.
[56] D. Triantafyllopoulou, N. Passas, A. K. Salkintzis, and A. Kaloxylos, "A heuristic cross-layer mechanism for real-time traffic over IEEE 802.16 networks," *Int. J. Netw. Manag.*, vol. 17, no. 5, pp. 347–361, 2007.
[57] G. Fairhurst, M. Berioli, and G. Renker, "Cross-layer control of adaptive coding and modulation for satellite Internet multimedia," *Int'l J. of Satellite Communications and Networking*, vol. 24, no. 6, pp. 471–491, 2006.
[58] I. Haratcherev, J. Taal, K. Langendoen, R. Lagendijk, and H. Sips, "Optimized video-streaming over 802.11 by cross-layer signaling," *IEEE Communications Magazine*, pp. 115–121, 2006.
[59] J. Makinen, P. Ojala, and H. Toukomaa, "Performance comparison of source controlled GSM AMR and SMV vocoders," *Proc. Int'l Symp. on Intelligent Signal Processing and Communication Systems*, pp. 151–154, 2004.
[60] T. Kawata and H. Yamada, "Adaptive multi-rate VoIP for IEEE 802.11 wireless networks with link adaptation function," *Proc. IEEE GlobeCom*, pp. 357–361, 2006.
[61] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streamig audio," *IEEE Network*, pp. 40–48, Sept.-Oct. 1998. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/c/CPOHVH98.pdf
[62] T. Z. J. Fu, D. M. Chiu, and J. C. S. Lui, "Performance metrics and configuration strategies for group network communication," in *IEEE Int'l Worksp. on Quality of Service*, 2007, pp. 173–181. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/t/TZJFDMCJCSL07.pdf
[63] D. Minoli and E. Minoli, *Delivering Voice over IP Networks*. New York, NY: Wiley Computer Pub., 1998.
[64] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RFC 1889: RTP: A transport protocol for real time applications," Jan. 1996, http://www.ietf.org/rfc/rfc3951.txt.
[65] H. Schulzrinne, "Real Time Protocol," 2008, http://www.cs.columbia.edu/~hgs/rtp/. [Online]. Available: http://www.cs.columbia.edu/~hgs/rtp/
[66] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "RFC 2205: Resource ReSerVation protocol (RSVP) – Version 1 functional specification," Sept. 1997, http://www.ietf.org/rfc/rfc2205.txt.
[67] P. S. Center, "Recent work on congestion control algorithms for non-TCP based applications," 2008, http://www.psc.edu/networking/projects/tcpfriendly/. [Online]. Available: http://www.psc.edu/networking/projects/tcpfriendly/
[68] S. Liang and D. Cheriton, "TCP-RTM: Using TCP for real-time multimedia applications," Dept. of Electrical Engineering, Stanford University, Stanford, CA, Tech. Rep., 2003. [Online]. Available: http://novel.crhc.illinois.edu/papers.db/l/LiaChe03.pdf
[69] P. J. Smith, P. Kabal, M. L. Blostein, and R. Rabipour, "Tandem-free VoIP conferencing: A bridge to next-generation networks," *IEEE Communications Magazine*, pp. 136–145, May 2003.
[70] Skype, "The Skype VoIP system," 2008, http://www.skype.com. [Online]. Available: http://www.skype.com
[71] QQTalk, "The QQ VoIP system," 2008, http://im.qq.com/qqtalk/help/create-system.shtml. [Online]. Available: http://im.qq.com/qqtalk/help/create-system.shtml
[72] M. Handley and J. Crowcroft, "Internet multicast today," *The Internet Protocol Journal*, vol. 2, no. 4, 1999.
[73] T. Montgomery, "Reliable multicast links," 2008, http://www.nard.net/~tmont/rm-links.html. [Online]. Available: http://www.nard.net/~tmont/rm-links.html
[74] I. R. W. Group, "Reliable Multicast Transport (RMT)," 2008, http://www.ietf.org/html.charters/rmt-charter.html. [Online]. Available: http://www.ietf.org/html.charters/rmt-charter.html
[75] D. Doval and D. O'Mahony, "Overlay networks: A scalable alternative for P2P," *IEEE Internet Computing*, pp. 1–5, 2003.

Benjamin W. Wah is currently the Franklin W. Woeltge Endowed Professor of Electrical and Computer Engineering and Professor of the Coordinated Science Laboratory of the University of Illinois at Urbana-Champaign, Urbana, IL. He also serves as Director of the Advanced Digital Sciences Center in Singapore, a research center between the University of Illinois, Urbana-Champaign, and the Agency for Science, Technology and Research (A*STAR). He received his Ph.D. degree in computer science from the University of California, Berkeley, CA, in 1979. Previously, he had served on the faculty of Purdue University (1979-85), as a Program Director at the National Science Foundation (1988-89), as Fujitsu Visiting Chair Professor of Intelligence Engineering, University of Tokyo (1992), and McKay Visiting Professor of Electrical Engineering and Computer Science, University of California, Berkeley (1994). In 1989, he was awarded a University Scholar of the University of Illinois; in 1998, he received the IEEE Computer Society Technical Achievement Award; in 2000, the IEEE Millennium Medal; in 2003, the Raymond T. Yeh Lifetime Achievement Award from the Society for Design and Process Science; in 2006, the IEEE Computer Society W. Wallace-McDowell Award and the Pan Wen-Yuan Outstanding Research Award, and in 2007, the IEEE Computer Society Richard E. Merwin Award and the IEEE-CS Technical Committee on Distributed Processing Outstanding Achievement Award. Wah's current research interests are in the areas of nonlinear search and optimization, multimedia signal processing, and computer networks.

Wah cofounded the IEEE Transactions on Knowledge and Data Engineering in 1988 and served as its Editor-in-Chief between 1993 and 1996, and is the Honorary Editor-in-Chief of Knowledge and Information Systems. He currently serves on the editorial boards of Information Sciences, International Journal on Artificial Intelligence Tools, Journal of VLSI Signal Processing, and World Wide Web. He had chaired a number of international conferences, including the 2000 IFIP World Congress and the 2006 IEEE/WIC/ACM International Conferences on Data Mining and Intelligent Agent Technology. He has served the IEEE Computer Society in various capacities, including Vice President for Publications (1998 and 1999) and President (2001). He is a Fellow of the AAAS, ACM, and IEEE.

Batu Sat is a Ph.D. candidate in the Department of Electrical and Computer Engineering at the University of Illinois, Urbana-Champaign. He received his M.S. degree in Electrical Engineering from Illinois in 2003 and his B.S. degree in Electronics and Telecommunications Engineering from Istanbul Technical University in 2001. His current research interests are in the development of evaluation and design methods for real-time multimedia communication systems. He's a student member of the IEEE and ACM.