QUALITY ASSESSMENT OF SPEECH CODECS IN SYNCHRONOUS E-LEARNING ENVIRONMENTS

Juan C. Granda, José Quiroga, Daniel F. García, Francisco J. Suárez

Department of Computer Science, University of Oviedo
Campus de Viesques, 33204, Gijón, Spain
Email: {jcgranda,uo81580,dfgarcia,fjsuarez}@uniovi.es

ABSTRACT

Synchronous e-learning tools always include an audioconference feature so participants in e-learning sessions can communicate orally. Audio quality is critical to avoid misunderstandings and improve user experience. Therefore, it is important to evaluate the audio quality that speech codecs can provide. Few existing assessment works consider the resources consumed (CPU or bitrate) to provide the audio quality, although synchronous e-learning sessions usually involve various participants, making resource consumption an important issue. In this paper, both objective and subjective audio quality measurement methods are used to characterize and estimate the audio quality of twenty speech codecs as a function of the resources consumed during synchronous e-learning sessions. Although users' opinions on audio quality are often more pessimistic than the evaluation provided by objective measurements, the correlation between objective and subjective measurements is high for medium-quality codecs. Users perceive lower audio quality for low-quality codecs than indicated by objective measurements, while they are not able to identify high-quality codecs, scoring them similarly to medium-quality codecs.

Index Terms— Speech quality, Mean Opinion Score, objective measurement, subjective measurement, synchronous e-learning

1. INTRODUCTION

Audioconference is one of the most useful features of synchronous e-learning tools. The audio channel conveys speech from the instructor and feedback from learners, so the instructor can adapt the pace of the class to learners' requirements.

Specific audio codecs are used to encode human voice in Voice over IP (VoIP) systems and synchronous e-learning sessions [1]. These codecs achieve high compression rates when compressing human voice, but they are not suitable for encoding other kinds of audio signals. Higher compression rates usually imply low audio quality and higher CPU resources consumed. The latter is especially important in synchronous e-learning sessions, where many participants may be interacting and also using other high CPU-consuming media such as video. Thus, audio quality, output bitrate and CPU resources consumed must be considered when selecting a speech codec for synchronous e-learning sessions.

Audio streams in synchronous e-learning sessions are usually longer than in VoIP conversations. In fact, the audio stream from the instructor carries educational information during the whole e-learning session. Audio hiss, background noise or audio glitches decrease audio quality, and the duration of synchronous e-learning sessions (1 to 2 hours) further lowers user quality perception. Although the quality provided by a codec may be considered appropriate for VoIP conversations, it may be considered low for synchronous e-learning sessions, as audio artifacts in the audio stream from the instructor are especially annoying for students attending lengthy sessions. Artifacts may be caused by the use of inappropriate hardware such as desktop microphones introducing impairments in the audio signals. Furthermore, many audio streams may be active simultaneously to support collaborative learning, making the identification of the speaker difficult.

The aim of this work is to characterize the quality provided by several speech codecs as a function of resources consumed in synchronous e-learning environments. Firstly, audio quality is estimated with objective measurements to classify them into audio quality categories. Then, subjective measurements involving users are carried out using recorded synchronous e-learning sessions.

The remainder of this paper is organized as follows. Technical background about speech codecs and audio quality assessment is presented in Section 2. In Section 3, related work on speech codec comparison is discussed thoroughly. The proposed codec evaluation is addressed in Section 4, and the results are exposed in Section 5. Finally, Section 6 presents the concluding remarks and outlines future work.

2. TECHNOLOGICAL BACKGROUND

This work focuses on speech codecs rather than wideband codecs, as the latter produce higher bitrates. Output bitrate

978-1-61284-350-6/11/$26.00

is important in synchronous e-learning sessions, especially in webinars with dozens of attendees. The lower the bitrate needed, the higher the number of audio streams which can be used simultaneously. The most used speech codecs in VoIP systems and synchronous e-learning platforms are examined: G.711, G.726 (bitrates 16 kbps, 24 kbps, 32 kbps and 40 kbps), G.729, iLBC (packet sizes 20 ms and 30 ms), Speex (modes from 0 to 10) and MELP.

2.1. Audio quality measurement

Several measurement methods are used for the estimation of audio quality. Many of them are analyzed in [2] and [3]. Measurement methods can be classified into two main categories: subjective and objective methods.

Subjective methods imply listening tests in which users are requested to score the audio quality of a number of samples. Audio quality is estimated as the average score. Although listening tests provide the best evaluation of the audio quality perceived by users, they require a high number of users to make the results statistically significant. Furthermore, a special testing room must be prepared for tests to be carried out in a quiet environment.

The most widely used subjective audio quality measurement methods are from the ITU-T: P.800 and all its variants (P.830, P.831, P.832 and P.835). These methods define procedures for the statistical calculation of audio quality from users' perceptions. They also include requirements for the environmental conditions of listening tests such as room characteristics and background noise levels. All of these subjective methods use Mean Opinion Score (MOS), defined in P.800.1, as their audio quality metric. MOS translates the users' subjective evaluations into scores in a range from 1 to 5 (from bad to excellent).

Objective methods use algorithms for estimating the audio quality of samples based on the audio signal, so subjectivity is avoided. These methods are less expensive to apply than subjective methods as neither listening tests involving users or preparing a testing environment are necessary. However, the results obtained with objective measurements do not always correlate with the audio quality perceived by users.

Objective measurement methods can be classified into two categories: perceptual and non-perceptual methods [3]. Non-perceptual methods are based on physical characteristics of signals, while perceptual methods are based on the operation of the human auditory system. Non-perceptual methods can be used as a first approach for the estimation of audio quality, although the correlation between measured and perceived audio quality may not be sufficiently accurate [2]. In general, speech quality estimations from perceptual methods are better than those from non-perceptual methods, since the latter are not especially focused on audio signals.

ITU-T has published a popular perceptual measurement method: Perceptual Evaluation of Speech Quality (PESQ). PESQ is specifically oriented to speech audio and has become a worldwide standard for assessing VoIP systems. It defines an algorithmic procedure for comparing the original voice signal with a degraded version. PESQ provides a score similar to the MOS obtained in listening tests.

Many authors propose new audio quality measures, which are usually compared to existing ones. PESQ is most often used as a reference, as it is clearly established as a good estimation of speech quality as perceived by users.

3. RELATED WORK

Several quality assessment works use subjective methods based on listening tests, in which listeners rate the quality perceived for different codecs. One of the first attempts, developed by the European Broadcasting Union (EBU) [4], evaluated 20 codecs. They obtained small confidence intervals and thus reliable and stable results, demonstrating that subjective evaluations can be repeatable and reproducible. A subsequent effort by the EBU evaluated 8 low-bitrate audio codecs [5]. Quality was evaluated for audio samples of 16 kbps mono and for a range from 20 to 64 kbps stereo. They concluded that scores obtained are dependent on the test audio samples used.

Light [6] compared 7 codecs using the MOS scores of listeners grouped according to their ages. They considered that results from different groups cannot be compared, as many factors affect the subjective rating of the quality of codecs.

Rämö and Toukomaa [7] compared 12 codecs in Nokia Labs using MOS scores, developing several experiments grouped in three sets. In the first set they compared two versions of AMR codec: narrowband against wideband. The goal of the second set was to compare narrowband codecs: proprietary against open-source. Similarly, they compared wideband codecs in the third set: proprietary against open-source. They concluded that wideband codecs perfom better than narrowband.

Due to the high cost of subjective methods for evaluating audio quality some researchers prefer to use objective methods. Hall [8] compared two codecs using four objective metrics and calculated the correlation of each metric with MOS.

Beuran and Ivanovici [9] obtained the PESQ scores of 4 VoIP codecs as a function of the average packet loss rate and the average jitter suffered by audio packets during transmission. Nguyen et al. [10] compared the degradation of audio quality of 9 VoIP codecs when the bit error rate was progressively increased. They expressed the quality using PESQ scores. Their evaluation was oriented to select the most robust codec for space communications.

Most of the comparative works discussed here do not provide results of the audio quality of codecs related with the consumed resources such as processor time or bitrate. Additionally, the available works use subjective or objective methods to evaluate quality, but rarely combine the two. Finally, most of the works discussed focus on VoIP but not on the spe-

cific characteristics of synchronous e-learning environments with many overlapped and lengthy audio streams. In this paper, an objective method to automate the experimentation (PESQ) is combined with subjective methods to validate the results, which are related to consumed resources in a synchronous e-learning environment.

4. QUALITY ASSESSMENT

Twenty speech codec configurations have been assessed to select the most suitable for synchronous e-learning sessions.

Codecs are analyzed according to audio quality and consumed resources. CPU encoding and decoding times and output bitrates are used to estimate the resources required by each codec, while objective audio quality measurements based on PESQ and subjective methods are used to evaluate their output audio streams during synchronous e-learning sessions.

There are so many audio codecs used to encode human voice that an exhaustive evaluation of all the available codecs would be prohibitively expensive, especially if subjective audio quality measurements involving users were applied. A filtering process is necessary to reduce the number of codecs to be assessed subjectively.

Initially, codecs are evaluated using objective measurements. CPU encoding and decoding times as well as output stream bitrates are measured for all codecs. Furthermore, audio quality is estimated using the PESQ method. The resulting PESQ scores allow for a classification of codecs in function of their audio quality and consumed resources.

Finally, a small subset of all codecs is selected from those analyzed to validate their objective audio quality estimations with subjective audio quality measurement methods during synchronous e-learning sessions.

4.1. Experimental environment

All the codecs have been implemented as DirectShow filters. DirectShow is a multimedia framework that handles multimedia data as streams flowing through chains of filters, known as filter graphs. Each filter performs an operation over the stream before passing data to the next filter in the chain.

The overload introduced by DirectShow is almost the same for each codec. Thus, the difference observed in the encoding and decoding time is due to the algorithmic complexity of the codecs. Although speech codecs are designed for real-time processing and operation in devices with limited processing capabilities (such as mobile devices), as users usually participate in synchronous e-learning sessions from desktop PCs, the processing capability needed to cope with the overload introduced by DirectShow is minimal.

The codecs have been integrated in the synchronous e-learning platform proposed in [11], so subjective audio quality tests can be carried out.

4.2. First phase: objective measurement

The first phase of the assessment process implies the estimation of the audio quality of each codec based on the PESQ scores and the measurement of the resources consumed. This is useful to propose an initial classification of the codecs.

The PESQ scores are calculated and the encoding and decoding times are measured for all the codec configurations. This is achieved using automated tests where audio samples are encoded and decoded, so times can be measured and PESQ scores can be computed comparing the original and decoded samples.

The encoding and decoding processes decrease the audio quality of the output audio samples. Audio quality degradation is measured with the PESQ method. The result of applying the PESQ method is an estimation of the subjective MOS, that is 4.5 for no degradation and a lower value when audio quality degradation appears.

Each codec was tested in this phase with 21 audio samples from public audio databases, including those from ITU-T and Carnegie Mellon University. All audio samples were encoded and decoded on the same computer equipped with an Intel Core 2 processor.

4.3. Second phase: subjective measurement

Once the objective PESQ scores are obtained, a subset of the codecs is selected to measure their audio quality subjectively with the opinion of users during synchronous e-learning sessions. Thus, the objective results can be validated.

The number of participants in a synchronous e-learning session varies as users join and leave the session. Usually, many participants interact at the same time, so many audio streams are interchanged among participants during the session. This implies high consumption of both network and processing resources. Thus, depending on the number of simultaneous audio streams and participants, high quality speech codecs are not always suitable for e-learning environments.

Three codec categories are considered in this work: low, medium and high audio quality. Codecs are classified into these categories according to objective audio quality measurements. Codecs from each of these categories are selected for subjective audio quality measurements. In order to select codecs, in addition to their audio quality, their bitrates and CPU times are considered. Usually a lower bitrate implies higher encoding and decoding times. However, the codec bitrate can be considered a more valuable resource than CPU, since computing power can be easily improved in desktop computers to reduce encoding and decoding times.

Subjective audio measurements require listening tests, in which users score the audio quality of a sequence of samples. These samples are encoded with the previously selected codecs. The users listen to the encoded samples and score their audio quality one at a time.

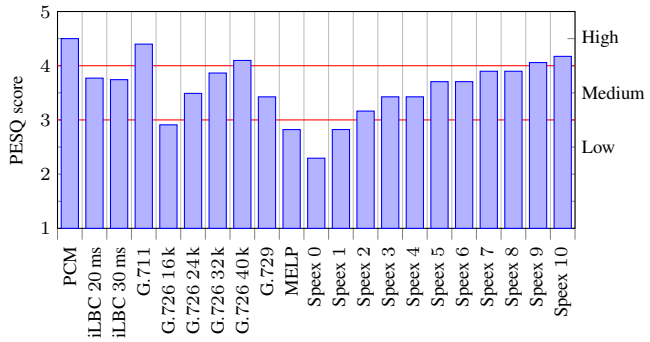

Fig. 1. Objective audio quality measurements

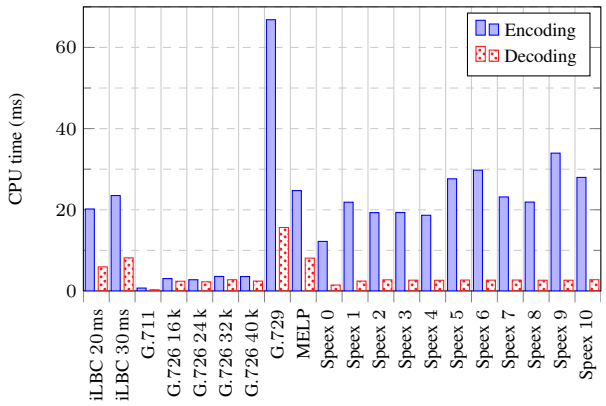

Fig. 2. CPU time for encoding and decoding 1 second of audio

Users should be asked to evaluate samples as representative as possible of synchronous e-learning sessions. Audio from synchronous e-learning has specific characteristics. Usually, audio streams from participants are captured from low-quality devices, such as desktop michrophones, so the original quality of the audio signals is relatively low. The encoding and decoding processes also decrease this quality. Furthermore, the instructor usually speaks for a long time without interruption, so samples must be long to emulate instructor speech. In occasions, many users speak at the same time when they are collaborating in a session.

In contrast to the previous phase, where audio samples were obtained from public databases, samples used to subjectively measure audio quality were directly captured from real synchronous e-learning sessions using desktop microphones. These sessions took place in an undergraduate course involving dozens of users and using the tool presented in [11]. Audio streams from the instructor and the students of an e-learning session were recorded and mixed in a single stream. This session was carried out within a local area network, so there was no decrease in audio quality due to network issues such as packet loss and interarrival jitter. The mixed audio stream was encoded with the selected codecs for the listening tests.

5. RESULTS

Fig. 1 shows the PESQ scores obtained for each codec after applying the objective audio quality PESQ method. Uncompressed PCM audio reaches the highest score (4.5), although its bitrate is extremely high (128 kbps). With the exception of G.711, which has a score close to PCM audio and half of its bitrate, the PESQ score is rarely greater than 4. G.726 40 kbps and Speex modes 9 and 10 obtain a high score, although Speex modes 9 and 10 need a slightly lower bitrate to achieve such high scores (18.2 and 24.6 kbps against 40 kbps). As shown in Fig. 1, codecs can be classified into low (score$\leq$3), medium (3$<$score$\leq$4), and high quality (score$>$4).

The results given by some codecs vary with the operation mode. G.726 obtains a high score operating at 40 kbps, but it decreases with the rest of the available modes. On the other hand, Speex obtains a higher score with significantly lower bitrate. In fact, Speex mode 7 obtains a score of 3.90, similar to the score obtained by G.726 32 kbps (3.86), despite producing an audio stream of lower than half the bitrate of G.726 (15 kbps against 32 kbps). iLBC obtains a similar score for its two modes, but iLBC 30 ms gets a lower bitrate. However, Speex obtains a higher score for the same bitrate. G.729 obtains a remarkable score of 3.43 given that it uses a bitrate of 8 kbps, but Speex modes 3 and 4 obtain a similar score using the same bitrate.

Fig. 2 shows the average CPU times required by each codec to encode and decode an audio sample with a duration of 1 second. All codecs have similar CPU times for encoding with the exceptions of G.729, G.711 and G.726. While G.729 is by far the most CPU-consuming codec, G.711 requires the minimum CPU time. Decoding times are almost identical for all the codecs except for G.729, MELP and iLBC, which are higher, and G.711 which is the lowest. Due to its low encoding times, G.726 can be used on devices with low processing capabilities or when it is necessary to manage many audio streams. On the other hand, decoding times for G.726 and Speex are constant for all their operation modes, while iLBC decoding times are different for its two modes. MELP and iLBC decoding times are significantly greater with respect to Speex and G.726.

Fig. 3 shows the PESQ score as a function of the output bitrate. In general, there is a direct relation between them. Two almost linear tendencies can be seen in the figure. The double dashed line shows the tendency of score and bitrate when changing the operation mode of G.726. The single dashed line shows the different operation modes of Speex and iLBC. This suggests that Speex and iLBC are better than G.726 because G.726 needs a higher bitrate to achieve a similar audio quality. As can be seen in the zoomed-in area, several codecs have similar score/bitrate relations, for example, Speex modes

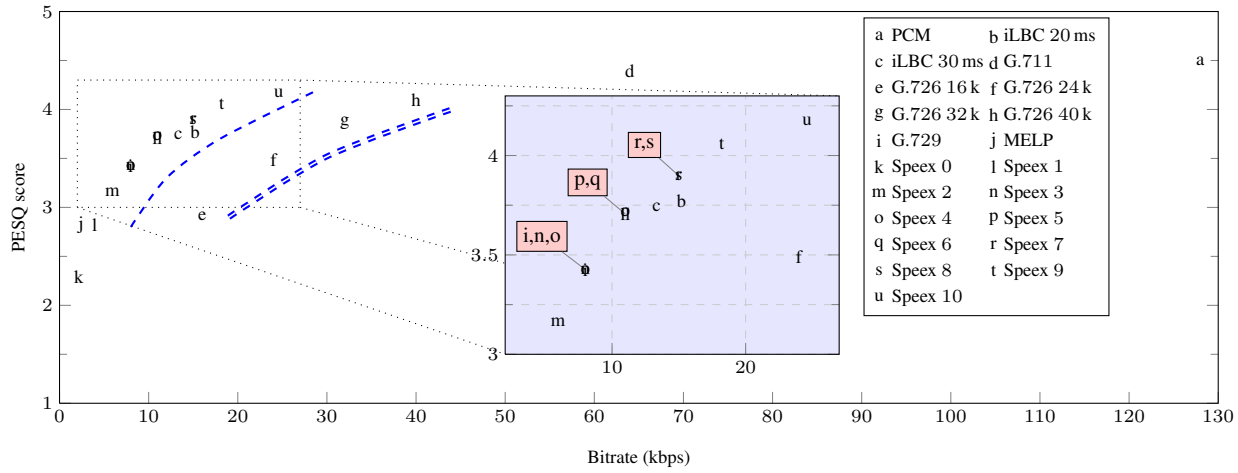

Fig. 3. Codec audio quality compared to bitrate

3 and 4, and G.729, although the computational complexity of the latter is significantly higher. Speex and iLBC use compression algorithms of similar complexity, although Speex provides slightly better results than iLBC.

The first evaluation phase has proved useful for the characterization of the codecs with a highly automated procedure using objective measurements. For the subjective user tests of the second evaluation phase, five codecs were selected, covering the previously established audio quality categories: low quality (Speex mode 1), medium quality (iLBC 30 ms, G.726 24 kbps and Speex mode 5) and high quality (Speex mode 9).

A room equipped with a computer and headphones, similar to those usually employed in synchronous e-learning sessions, was prepared for listening tests. In this way, users participated in the listening tests in a quiet environment.

The tests involved 44 users, both laymen and experts in audio encoding. All the users were under 35 and were familiar with information technologies. This is the user profile to which synchronous e-learning sessions are usually oriented. They listened to the audio samples and scored them in the testing room. Fig. 4 shows the subjective MOS obtained from these experiments compared to the objective score previously obtained. Confidence error bars for a confidence level of 95% are also shown in the figure.

The comparison between subjective and objective MOS indicates that users perceive little difference between iLBC 30 ms, G.726 24 kbps and Speex operating in modes 5 and 9. The audio quality of all of these codecs is perceived by users as good (subjective MOS slightly higher than 3.5), which corresponds to an objective MOS in the range of 3.0 to 4.0. Audio quality perceived by users for Speex 1 is poor, although its objective MOS is higher.

The error bars in Fig. 4 show that objective MOS for medium-quality codecs (iLBC 30 ms, G.726 24 kbps and Speex mode 5) obtained using the PESQ method and subjective MOS can be considered equivalent, but not for low-quality (Speex mode 1) and high-quality codecs (Speex mode 9). This means that users are not able to perceive high-quality codecs. It is likely that background noise in the original audio signal due to the low-quality devices used to capture audio (usual in synchronous e-learning) prevents users from appreciating high-quality codecs. This background noise is even more noticeable with low-quality codecs, so users scored them lower.

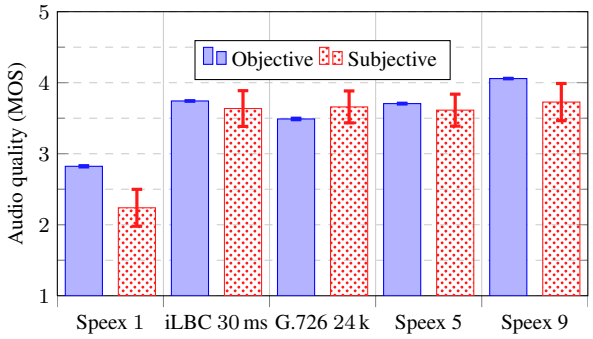

Fig. 4. Comparison of MOS: subjective against objective

Users were requested to answer some questions after the listening tests. The first question evaluated the effort that users needed to make in order to listen to the speech of the speakers in the audio samples. Fig. 5 shows the responses of users. Approximately 90 % of users needed minimal or no effort to listen to the speech for medium-quality and high-quality codecs, while only 34.1 % of users were able to listen to the speech easily for Speex mode 1. It must also be noted that, despite Speex mode 9 scoring higher than G.726 24 kbps in Fig. 4, a higher number of users reported needing no effort in order to listen to the speech for the latter (54.55 %) than for the former (52.27 %).

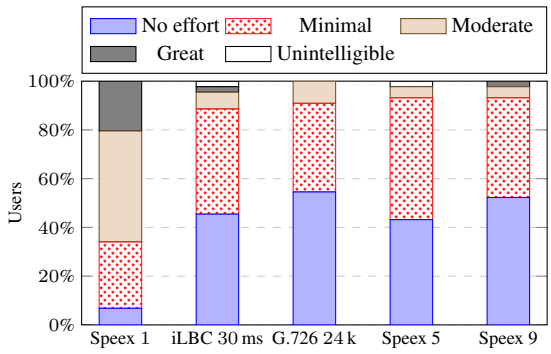

Fig. 5. Effort made by users to listen to the speech

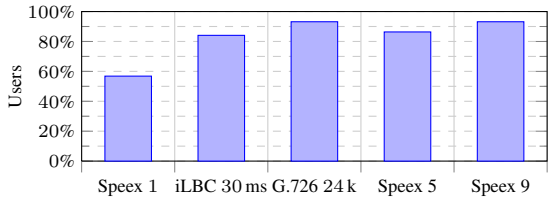

Fig. 6. Users that would be able to identify the speaker

Fig. 6 shows the opinion of users as to whether they can identify or would be able to identify the speaker in the audio samples. Again, G.726 24 kbps and Speex mode 9 obtain a similar result: almost all the users would be able to identify the speaker. The identification of the speaker is impossible in audio samples encoded with Speex mode 1 for 43.18 % of users, while 79.55 % of users are able to listen to the audio speech with a moderate effort.

6. CONCLUSIONS

An in-depth assessment of audio quality and resources consumed by speech codecs has been presented. Both objective and subjective measurements have been used, the former to classify codecs into three audio quality groups, and the latter to validate results. The evaluation process concludes that the correlation between objective and subjective audio quality measurements is high for medium-quality codecs, so both objective and subjective methods may be used interchangeably.

However, users cannot perceive audio quality differences between medium-quality codecs and high-quality codecs, and they perceive lower quality for low-quality codecs than that reflected by objective measurements.

Medium-quality codecs, MOS $\in (3, 4)$, are the most suitable for synchronous e-learning, as they provide a similar subjective audio quality to high-quality codecs while using fewer resources. Furthermore, users can easily listen to audio conversations and speakers when audio is encoded with medium-quality or high-quality codecs. Additionally, Speex mode 5 is the medium-quality codec with the best compromise between computational complexity and output bitrate.

7. REFERENCES

[1] D. F. García, C. Uría, J. C. Granda, F. J. Suárez, and F. González, "A functional evaluation of the commercial platforms and tools for synchronous distance e-learning," *Int. J. of Educ. and Inf. Technol.*, vol. 1, no. 2, pp. 95–104, 2007.

[2] K.H. Lam, O.C. Au, C.C. Chan, K.F. Hui, and S.F. Lau, "Objective speech quality measure for cellular phone," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process. (ICASSP'96)*, May 1996, vol. 1, pp. 487–490.

[3] Hamza Özer, Ismail Avcibaş, Bulent Sankur, and Nasir D. Memon, "Steganalysis of audio based on audio quality metrics," in *Proc. SPIE Electron. Imaging Conf. on Secur. and Watermarking of Multimed. Contents*, Jan. 2003, vol. 5020, pp. 55–66.

[4] EBU, "Listening tests on Internet audio codecs," Technical Review 283 of EBU (European Broadcasting Union), June 2000.

[5] EBU, "Subjective listening tests on low-bitrate audio codecs," Technical Review 3296 of EBU (European Broadcasting Union), June 2003.

[6] J. Light and A. Bhuvaneshwari, "Performance analysis of audio codecs over real-time transmission protocol (RTP) for voice services over Internet protocol," in *Proc. of 2nd Annu. Conf. on Commun. Netw. and Serv. Res. (CNSR'04)*, May 2004, pp. 351–356.

[7] A. Rämö and H. Toukomaa, "On comparing speech quality of various narrow- and wideband speech codecs," in *Proc. of 8th Int. Symp. on Signal Process. and its Appl. (ISSPA'05)*, Aug. 2005, vol. 2, pp. 603–606.

[8] T. A. Hall, "Objective speech quality measures for Internet telephony," in *Proc. SPIE Conf. on Voice Over IP (VoIP) Technol.*, Aug. 2001, vol. 4522, pp. 128–136.

[9] R. Beuran and M. Ivanovici, "User-perceived quality assessment for VoIP applications," Tech. Rep. CERN-OPEN-2004-007, CERN, Geneve, Switzerland, Jan. 2004.

[10] S. Nguyen, C. Okino, L. Clare, and W. Walsh, "Space-based voice over ip networks," in *Proc. IEEE Aerosp. Conf. (AC'07)*, Mar. 2007, pp. 1–11.

[11] J. C. Granda, D. F. García, F. J. Suárez, I. Peteira, and C. Uría, "A multimedia tool for synchronous distance e-training of employees in geographically dispersed industries," in *Proc. IASTED Int. Conf. on Web-based Educ.*, Mar. 2008, pp. 1–8.