On the Language and Gender Dependency of Instrumental Perceptual Speech Quality Evaluation in Internet-Telephony

Florian Hammer
Telecommunications Research Center Vienna (ftw.)
Donau-City-Strasse 1
A-1220 Vienna, Austria
hammer@ftw.at

Ian Marsh
Swedish Institute of Computer Science (SICS)
Box 1263, Kista 164-29,
Sweden
ianm@sics.se

Thomas Ziegler
Telecommunications Research Center Vienna (ftw.)
Donau-City-Strasse 1
A-1220 Vienna, Austria
ziegler@ftw.at

ABSTRACT

The successful deployment of VoIP requires an acceptable of perceptual speech quality. In this paper, we assess the language and the gender dependency of PESQ, a standard algorithm for instrumental speech quality evaluation, by employing traces gathered from wide area Internet measurements. We use the iLBC codec and speech samples in French, Japanese and English and an "Artificial voice" which reproduces the characteristics of the human voice. We observe that PESQ overestimates the quality of French, Japanese, and the Artificial voice speech samples as compared to English, the language PESQ was originally designed for. Additionally, we observe a bias against female speakers.

In order to verify the results which are based on wide-area network measurements, we repeat the investigations using artificially generated trace fragments. As to ensure there is no bias towards a particular language or gender due to artifacts introduced by the low-bitrate speech coding, we use the G.711 codec in the generalized study. Our observations show that PESQ exhibits a positive bias for non-English languages and male speakers. Therefore, we conclude that the applicability of PESQ is limited in the context of Voice over IP quality assessment.

Keywords

Gender dependency, language dependency, PESQ, VoIP

1. INTRODUCTION

Voice over IP (VoIP) has emerged as a widespread application in the Internet partly due to recent advances in speech coding especially designed for networks offering best effort service as well as the upcoming of peer to peer applications like Skype. What is important is the quality of a voice conversation perceived by the end user. When evaluating the performance of VoIP, however, Internet engineers have stuck to classical network centric QoS metrics like delay, delay jitter, and loss probability to evaluate the quality of a voice conversation. Only few investigations exist aiming at understanding how the variation of QoS parameters in the underlying transport network affect the quality of a speech conversation as perceived by the user.

In this paper, we contribute to the research task of perceptual speech quality evaluation in an IP based environment by performing perceptual speech quality evaluations using the PESQ algorithm [11] and VoIP traces gathered from wide area Internet measurements plus artificially generated traces. Using state of the art codecs our objective is to gain insights into the language and gender dependency of perceived speech quality as estimated by the PESQ algorithm. By eliminating the influence of low-bitrate speech coding on the quality and testing at a wide range of network conditions, we focus on the impact of the instrumental quality evaluation algorithm itself. Thus, our work contributes to an understanding of the applicability of such algorithms.

The paper is structured as follows. In section 2 we discuss related work about existing research on this subject. In section 3 we present the method we have used to evaluate the wide area packet traces. Section 4 presents results of the quality assessment of the global traces and interim conclusions drawn from them. In Section 5, we analyze the language and gender dependency on a general basis using artificially generated packet traces. Finally, section 6 concludes this paper.

2. RELATED WORK

This section presents work related to the PESQ algorithm and its use for investigating the impact of language and gender on the perceptual speech quality.

In 2001, ITU-T Rec. P.862 Perceptual Evaluation of Speech Quality (PESQ) has been standardized as instrumental evaluation tool for user perceived speech quality [11]. Figure 1 shows the structure of PESQ. A reference speech signal is transmitted through a network which results in a quality degradation corresponding to the network conditions and coding scheme. PESQ analyzes both the reference and degraded signal and calculates their representation in the perceptual domain based on a psychoacoustic model of the hu-

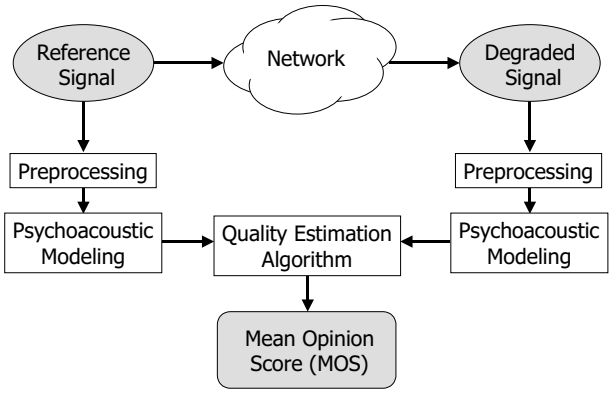

Figure 1: Structure of the PESQ Algorithm.

man auditory system. In this way, a disturbance between the original and the degraded speech signal can be calculated and the corresponding subjective Mean Opinion Score (MOS) can be estimated. Note that this estimation is based on the data resulting from subjective listening tests at various network conditions. The PESQ-MOS usually ranges from 4.5 (excellent perceived quality) down to 1 (bad perceived quality). PESQ has recently been used for predicting the perceptual quality of VoIP [6, 18].

Rix, one of the main developers of PESQ, has introduced a mapping function (PESQ-Listening Quality, PESQ-LQ) which better models the subjective mean opinion score (MOS), especially at low quality conditions [15]. Concerning the estimation of the speech quality in different languages, the estimation parameters are mainly driven by speech material in English. Rix concludes that French samples tend to be underestimated on average and Japanese samples are overestimated. Note that the data sets, and thus the variety of conditions for French and Japanese samples are limited compared to English samples. However, the mean correlation of PESQ-LQ for VoIP datasets was 0.933 [15].

Sun [17] presents an investigation of the relation between PESQ-evaluated perceived speech quality, gender and language for four different CELP-based (Code Excited Linear Prediction) speech codecs under various network conditions. Her study shows that the quality of female talkers is rated worse than that of male talkers with the same network conditions. Furthermore, the quality depends on the language, e.g. English being rated better than that of French. Sun states that the gender and language dependency is likely to be due to the speech coding algorithms referring to [2].

Mohamed et al. [14] have reported an influence of the languages Spanish, Arabic, and French on subjective quality perception using G.711 PCM speech coding and the insertion of silence as packet loss concealment. In many conditions, French samples performed best, however, no conclusions were drawn about the cause of this dependency.

In a study about the usefulness of corrupted speech data [5], we have experienced gender dependency for a set of German speech samples when using PESQ, but only marginal dependency when using another instrumental speech quality assessment algorithm (TOSQA, [7]). At that time, we concluded that the difference between PESQ and TOSQA may have been due to two reasons: One being the speech coding algorithm (in accordance with Sun's conclusion), and the other was that the instrumental assessment methods behave in different ways. However, the topic has not been investigated in more detail.

As an overall observation on related work and as a motivation for this paper we find that research on language and gender dependency of speech quality over the Internet is still limited and that contradictous statements exist. As examples we mention the findings on language dependency in [15], [17] and [14] and the results on the cause for language and gender dependencies in [17] and [5].

3. MEASUREMENTS

This section describes the data and evaluation methods we use to investigate the language and gender dependency. We first used measurement data from global VoIP measurements.

3.1 Wide Area Measurements of VoIP

This work is based on packet loss traces measured in VoIP wide area measurements carried out at SICS/KTH in 2003 [13]. To measure the network VoIP quality they sent prerecorded voice calls between globally distributed sites which are shown in Fig. 2. The intervening network paths were probed by a 70 second PCM coded pre-recorded call. The call consisted of 2043/3652 packets with 160 byte payloads being transmitted with and without silence suppression respectively. The goal of this measurement phase was investigate the network quality and to provide loss, delay and jitter for the sites chosen.

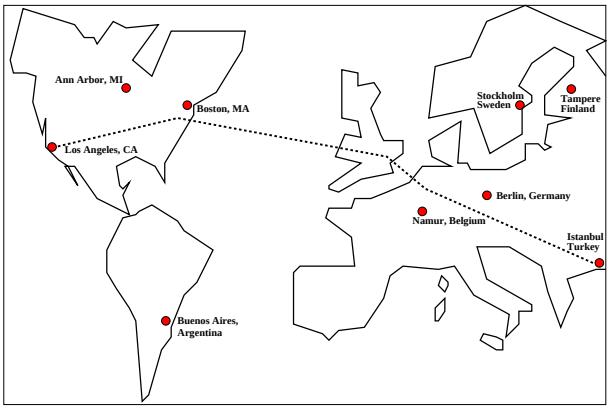

Figure 2: Global test sites with used ones marked.

Most of the connections between the distributed sites exhibited a negligible packet loss probability. Thus, their evaluation is of minor interest and we concentrate on 71 traces of the link between Los Angeles, California, USA, and Istanbul, Turkey. During the measurements we observed 23 hops from California to Turkey as reported by traceroute. The

mean packet loss from the Turkish host to the Californian was 7.6% (Standard deviation: 6.8%) and from the Californian host to the Turkish one was 4.3% (2.4%). The mean delays were 410 ms (32 ms) from the Turkish site and 419 ms (25 ms). Finally the jitter was 8.8 ms (2.5 ms) from the Turkish host and 5.3 ms (1.7 ms) as calculated according to the RFC 3550 [16]. Note that the Turkish academic network is connected to the Internet in Europe via a satellite link using the Geant network.

3.2 Evaluation Method

In a VoIP scenario, the perceptual speech quality heavily depends on the amount of packet loss which occurs within the network. Since we focus on "listening-only" quality testing (no interaction) and PESQ is not able to estimate the quality impairment caused by transmission delay, in these experiments we can neglect the influence of the delay on the quality.

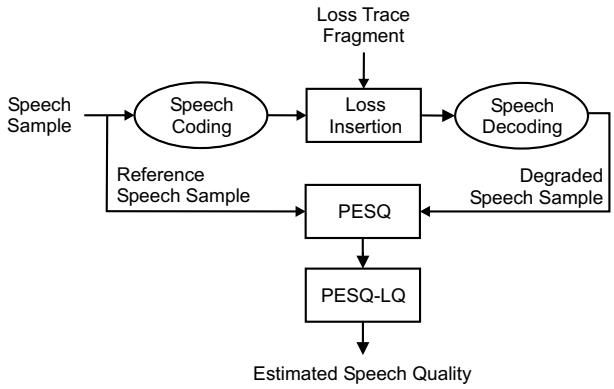

Figure 3: Evaluation Method. The loss pattern of a trace fragment is applied to the bitstream of a coded speech sample. After decoding the degraded bitstream, the quality of the resulting speech sample is estimated using the PESQ algorithm and the PESQ-LQ mapping function.

The measurement method is shown in Figure 3. We consecutively extract fragments of the size of a speech sample from a packet trace and apply the packet loss pattern to the bitstream of the coded speech sample. The quality of the degraded speech sample is then evaluated by PESQ. As a refinement of the PESQ evaluation results, we use a mapping function PESQ-LQ as presented in Section 2.

As a speech coding algorithm we chose the Internet Low-Bitrate Codec (iLBC) which was standardized within the IETF in RFC 3951 [1]. The iLBC codec is especially designed for voice transmission in a packet-based network and has been shown to be robust under conditions of high packet loss. It codes speech sampled at 8 kHz/16 bits using a frame size of 20 ms which results in a bitrate of 15.2 kbps.

The speech database used in this study includes a set of 16 speech samples (2 female and 2 male speakers) for each of the following three languages: English, French, and Japanese. These speech samples are taken from the ITU-T speech database [9]. Each speech sample is 8 seconds long corresponding to 400 packets at a frame size of 20 ms. In addition, we use two artificial speech samples (one female and one male) which are taken from ITU-T P.50 Appendix 1 [8]. Artificial voice, as described in [10], reproduces the time and spectral characteristics of speech and has previously been used for VoIP quality measurement [3, 4]. The artificial voice samples are 11.1 seconds long corresponding to 555 packets per speech sample at a frame size of 20 ms.

4. MEASURED TRACE RESULTS

This section presents the results from the perceptual speech quality evaluation of the wide area measurement data as outlined in section 3.1. The evaluation results for the connection Los Angeles-Istanbul is shown in Figure 4. In general, we can clearly observe the degradation of the perceptual speech quality with increasing packet loss rate. As the loss probability per fragment varies from 0 to 0.15 the PMOS-LQ degrades approximately from 4.2 to 2 for artificial voice, from 4.2 to 2.4 for French samples, and from 4 to 2.2 for Japanese samples, and from 3.7 to 1.5 for English samples. We observe that artificial voice and French speech samples are rated approximately the same. English samples are rated lowest and Japanese samples are in between.

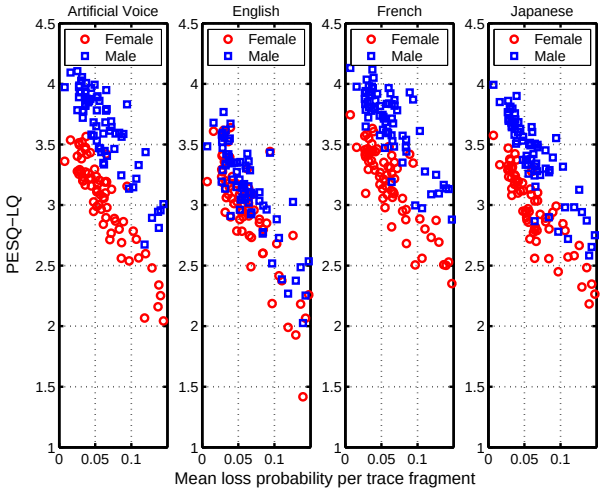

Figure 4: Results of the evaluation of the connection from Los Angeles to Istanbul. The speech samples were coded using the iLBC. Circles depict the PMOS-LQ for female speech samples and squares represent the results for male speech samples.

If we now distinguish between female and male speech samples, we can observe a clear difference among the genders for French and Artificial speech samples. Male samples are rated up to 1 PMOS-LQ better than female samples under the same loss conditions. For English speech samples, the gender difference is significantly less pronounced. Again, Japanese is in between.

We have experienced both language and gender dependency for the global traces evaluated using the iLBC codec. Now the question arises whether these differences result from the a) specific connection (Los Angeles to Istanbul), b) from the speech coding algorithm, or c) from the evaluation algorithm

(PESQ). This question will be addressed in the following section by testing artificially generated packet loss test traces.

5. COMPLEMENTARY INVESTIGATION

The previous section has shown that the speech quality evaluation of real packet loss traces results in different performance depending on language and gender. That study was carried out using a representative global VoIP connection, iLBC, and the PESQ algorithm. In order to perform a controlled study of the before mentioned dependencies, we use the G.711 codec as to avoid low-bitrate speech compression as a potential parameter of influence. To eliminate the potential influence of specific packet loss characteristics of the measured connection, we generate artificial trace fragments.

5.1 Trace Fragments

For the generation of trace fragments matching the size of the speech samples, we used the Gilbert Model ([12], see Figure 5) which is widely used for packet loss modelling. In this 2-state Markov model state 0 represents the successful reception of a packet, whereas state 1 signifies the loss of a packet. The transition probabilities are defined as follows: p is the probability that a packet will be lost given that the previous packet has been received, q is the probability that a packet will be received given that the previous packet has been lost. 1-q is termed "conditional loss probability" (CLP), which represents the probability that a packet is lost given that the previous packet was lost. Thus, the CLP serves as an indicator for the loss burstiness of the traffic.

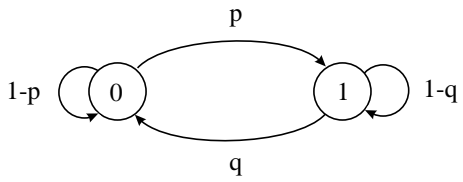

Figure 5: Gilbert Model.

Moreover, the "unconditional loss probability" (ULP) represents the average packet loss rate and can be calculated [12] as

$$ULP = \frac{p}{p+q}. \quad (1)$$

Due to the speech sample length as given by the ITU-T speech database, we have chosen trace fragment length of 400 packets for the real speech samples. The same applies to the artificial voice samples which are 11.1 seconds long which leads to a fragment length of 555 packets. The ULP of the traces we chosen in 5% steps from 0-20%, and the CLP={10,40,70}%. These conditions span a wide range of controlled loss situations. We have generated 30 different trace fragments per combination of ULP and CLP for statistical confidence.

As mentioned above, these traces were evaluated with the G.711 codec using its internal packet loss concealment at a frame size of 20 ms in order to get results comparable to those of the real traces.

5.2 Perceptual Quality

The results of the evaluation of the loss trace fragments generated by the Gilbert model are shown in Figures 6-8. The PESQ-LQ scores for the male speakers are shown on the left side and the scores for the female speakers are presented on the right.

Figure 6 show the perceptual quality estimates over the ULP for the case of low burstiness (CLP=10%). Here, we can observe distinct differences in the languages for male speakers which become more obvious with increasing average packet loss rate. At a ULP of 20%, the difference between English and French is 0.77 PESQ-LQ MOS points, while the difference between English and Japanese is 0.28 MOS points. As opposed to male talkers, the quality ratings for female talkers are consistent over languages with the exception of French. French is rated 0.5 MOS points higher than English at an average packet loss rate of 20%.

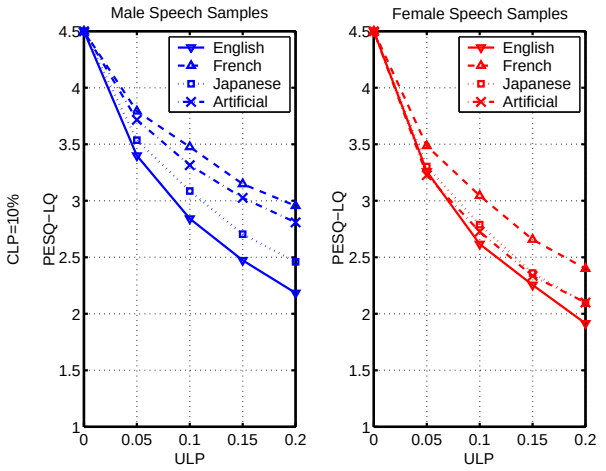

Figure 6: Mean PESQ-LQ values vs. ULP at CLPs of 10 %. Separated by gender. (Speech codec: G.711.)

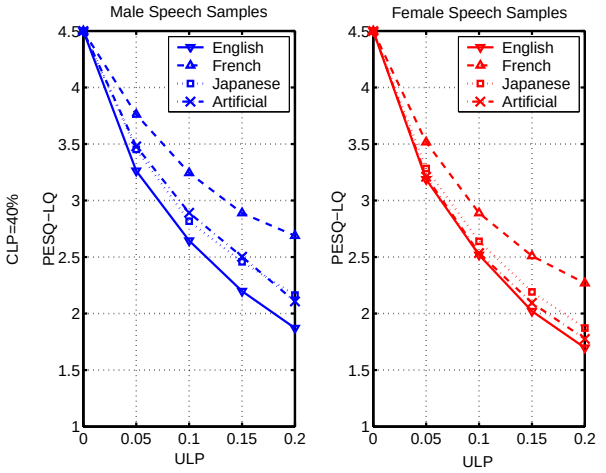

Figure 7: Mean PESQ-LQ values vs. ULP at a CLP of 40%. Separated by gender. (Speech codec: G.711.)

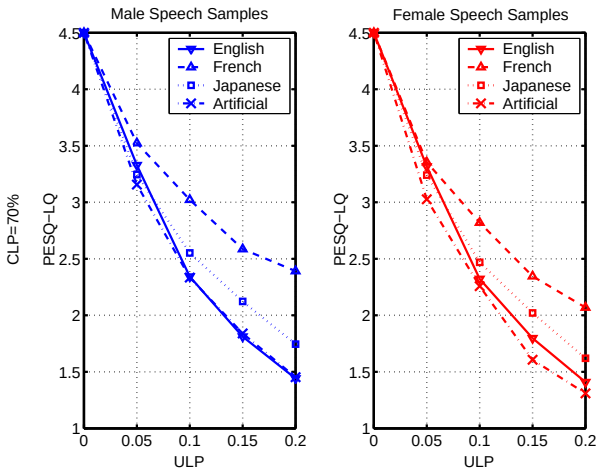

Figure 8: Mean PESQ-LQ values vs. ULP at a CLP of 70%. Separated by gender. (Speech codec: G.711.)

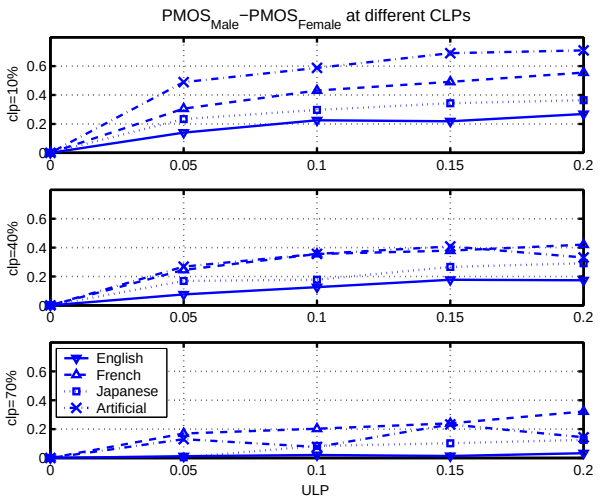

Figure 9: Difference between PESQ-LQ values for Male and Female speakers for four languages. (Speech codec: G.711.)

Increasing the burstiness to a CLP of 40% and 70%, respectively, we expectedly observe a decrease of PESQ-LQ MOS scores as compared to low burstiness. In case of a CLP=70% and ULP=20%, English male speech samples are rated with a MOS value of 1.5 as compared to 2.2 for a CLP of 10% in Figure 6.

For Female speech samples, we find that the ranking of the languages with respect to the speech quality does not depend on the CLP. On the contrary, for male speech samples, Artificial voice is close to French at CLP of 10%, and decreases to MOS values of English speech samples for a CLP of 70%. In order to illustrate the meaning of differences in MOS values, we consider the acceptable ULPs for a equal quality levels (same PESQ-LQ). For example, considering male users at a CLP of 10% in Figure 6, French users may have a ULP of about 20% wheras English users may have a ULP of 8% for an equal quality level of 3 MOS points.

To study the gender dependency, we calculate the difference between the PESQ-LQ scores of male and female talkers as depicted in Figure 9. This representation better illustrates the gender difference at all loss conditions for all languages under study. At a CLP of 10%, we can observe a clear increase in the difference between male and female speakers for French, Japanese and Artificial voice. The difference is largest for artificial voice, ranging from about 0.5 at a ULP of 5% up to 0.7 at a ULP of 20%, followed by French samples increasing from 0.3 up to 0.56 PESQ-LQ points at a ULP of 5% and 20%, respecively. As the CLP increases, the gender differences get reduced.

In general, we observe similar language and gender dependencies for the trace fragments with G.711 speech coding as shown for the measured traces in Section 4 using iLBC. Artificial voice, French, and Japanese samples are rated significantly better than English samples. Additionally, the quality of male speech samples is rated higher than female samples under comparable loss conditions. Consequently, as codec and loss traces are different in the evaluations in Sections 4 and 5, we conclude that PESQ introduces a considerable amount of bias against female talkers and for non-English languages.

6. CONCLUSIONS

In this paper, we have presented insights about the dependency of instrumental speech quality evaluation on language and gender of the talker. Using traces from a wide area Internet connection and applying iLBC speech coding, we observed significant dependencies of language and gender on the PESQ-estimated speech quality. In order to avoid potential influence of the low-bitrate speech coding and possible specific loss characteristics of the trace's Internet path, we have repeated the investigation using PCM coding and synthetic trace fragments. The results from this controlled study show that both gender and language dependency result from the PESQ algorithm. Those dependencies can be summarized as follows: The speech quality of French, Japanese and Artificial voice is overestimated as compared to English, the language PESQ is mainly based on. With regard to gender dependency, we found that the quality of male speech samples is rated higher than female samples under comparable loss conditions. Those observations are artifacts introduced by the PESQ algorithm, hence limiting its applicability in the context of VoIP quality assessment.

7. ACKNOWLEDGEMENTS

This work has been partly funded under the Austrian government's Kplus Competence Center Program and partly supported by the European Union under the E-Next Project FP6-506869.

8. REFERENCES

[1] S. Andersen, A. Duric, H. Astrom, R. Hagen, W. Kleijn, and J. Linden. Internet low bit rate codec (iLBC). *Request for Comments (Standards Track) RFC 3951, Internet Engineering Task Force*, Dec. 2004.

[2] P. A. Barrent, R. M. Voelcker, and A. V. Lewis. Speech transmission over digital mobile radio channels. *BT Technology Journal*, 14:4556, Jan. 1996.

[3] A. E. Conway. A passive method for monitoring voice-over-IP call quality with ITU-T objective speech quality mesurement methods. In *Proc. IEEE Int. Conf. Communications*, volume 4, pages 2583–2586, New York City, NY, 2002.

[4] H. Furuya, S. Nomoto, H. Yamada, N. Fukumoto, and F. Sugaya. Experimental investigation of the relationship between IP network performances and speech quality of VoIP. In *Int. Conf. Telecommunications*, volume 1, pages 543–552, Feb. 2003.

[5] F. Hammer, P. Reichl, T. Nordström, and G. Kubin. Corrupted speech data considered useful: Improving perceived speech quality of voip over error-prone channels. *Acta Acustica united with Acustica*, 90(6):1052–1060, Nov/Dec 2004.

[6] C. Hoene, S. Wiethlter, and A. Wolisz. Predicting the perceptual service quality using a trace of VoIP packets. In *Fifth International Workshop on Quality of future Internet Services (QofIS)*, Barcelona, Spain, Sept. 2004.

[7] International Telecommunication Union. TOSQA - Telecommunication objective speech quality assessment. *ITU-T COM 12-34*, Dec. 1997.

[8] International Telecommunication Union. Artificial voice, appendix I: Test signals. *ITU-T Recommendation P.50, Appendix I*, Feb. 1998.

[9] International Telecommunication Union. ITU-T coded-speech database. *ITU-T Series P, Supplement 23*, Feb. 1998.

[10] International Telecommunication Union. Artificial voice. *ITU-T Recommendation P.50*, Sept. 1999.

[11] International Telecommunication Union. Perceptual evaluation of speech quality (PESQ) , an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *ITU-T Recommendation P.862*, Feb. 2001.

[12] W. Jiang and H. Schulzrinne. Modeling of packet loss and delay and their effect on real-time multimedia service quality. In *Proc. Int. Workshop Network and Operating Systems Support for Digital Audio and Video NOSSDAV*, Chapel Hill, NC, June 2000.

[13] I. Marsh, F. Li, and G. Karlsson. Wide area measurements of voice over IP quality. In *Proc. QoFIS'03*, Stockholm, Sweden, Oct. 2003.

[14] S. Mohamed, F. Cervantes-Pérez, and H. Afifi. Integrating networks measurements and speech quality subjective scores for control purposes. In *IEEE INFOCOM01*, volume 2, page 641649, Anchorage, Alaska, Apr. 2001.

[15] A. W. Rix. Comparison between subjective listening quality and P.862 PESQ score. In *Proc. Measurement of Speech and Audio Quality in Networks (MESAQIN'03)*, Prague, Czech Republic, May 2003.

[16] H. Schulzrinne et al. RTP: A transport protocol for real-time applications. *Request for Comments (Standards Track) RFC 3550, Internet Engineering Task Force*, July 2003.

[17] L. Sun. *Speech Quality Prediction for Voice over Internet Protocol Networks*. PhD thesis, University of Plymouth, Jan. 2004.

[18] L. Sun and E. C. Ifeachor. New models for perceived voice quality prediction and their applications in playout buffer optimization for VoIP networks. In *Proc. IEEE Int. Conf. Communications*, Paris, France, June 2004.