WHEN SHOULD A SPEECH CODING QUALITY INCREASE BE ALLOWED WITHIN A TALK-SPURT?

Stephen D. Voran and Andrew A. Catellier

Institute for Telecommunication Sciences
325 Broadway, Boulder, Colorado, USA
{svoran,acatellier}@its.bldrdoc.gov

ABSTRACT

The value or harm associated with an increase in speech coding quality depends on the type of the increase as well as the temporal location of the increase in an utterance. For example, some increases in speech coding bandwidth can be perceived as impairments. The higher quality associated with the wider bandwidth can offset the impairment, but only if the increase happens early enough in an utterance. We present a subjective speech-quality experiment that qualifies these relationships at the talk-spurt time-scale for six different combinations of AMR and SILK speech coders. If a quality increase does not include a bandwidth increase, then, on average, it is beneficial only if it occurs in the first 2.8 seconds of a talk-spurt. If a quality increase includes a bandwidth increase, then it is beneficial only if it occurs in the first 1.8 seconds of a talk-spurt.

Index Terms— AMR, SILK, speech bandwidth, speech coding, speech quality, subjective testing, time-varying speech quality

1. BACKGROUND AND MOTIVATION

Available resources on modern voice networks vary with time. This, along with the mobility of many voice network users, results in dynamic resource availability for any given call. Service providers strive to provide a graceful degradation of speech quality when network resources become scarce during a call. When additional network resources become available during a call, it may be possible to increase the speech coding rate and deliver higher speech quality.

But the effect of the quality *transition* must be considered. For example, wideband (WB) speech (50 to 7000 Hz nominal passband) has a documented higher perceived quality than narrowband (NB) speech (300 to 3400 Hz nominal passband) [1]–[3], but a transition from NB to WB speech coding is perceived as an impairment [4]–[6]. If the transition happens early enough in a speech recording, the value of the WB portion can exceed the harm of the transition, for a net improvement (relative to NB only) in overall speech quality. This was the case for NB-to-WB transitions at the 15 or 30 second point in a 60 second recording [4], [5]. But if the transition happens later in a speech recording, the shorter duration of the WB portion means that its value does not overcome the harm of the transition. This was the case for NB-to-WB transitions at the 45 second point in a 60 second recording [4], [5] or at the three-second point of a six-second recording [6]. In [6] we also experimented with gradual transitions (up to 2.5 seconds long) but found they did not mitigate the harm of the transition.

Even quality transitions within a fixed bandwidth can be perceived as impairments. In [7], [8] short NB recordings with distinct quality levels were concatenated to form longer recordings and subjective scores were provided for both the short and long recordings. Analysis of these scores shows that when average quality is held constant, increases in quality variation lead to reductions in long-term speech quality.

In [9] subjects evaluated three-second NB speech recordings with a low-high-low quality profile. The insertion of the high-quality segment was judged to be a benefit, in spite of the fact that doing so required two transitions.

Thus it is clear that the value (or harm) associated with an increase in speech quality will depend on the type of the increase (especially bandwidth increases, if any) as well as the temporal location of the increase relative to the start and end of the utterance.

One could constrain increases in speech coding bit-rate to occur only between calls or between talk-spurts. But the least-constrained and most fundamental question is this: given that a talk-spurt is already in progress and additional network resources have become available, what speech coding improvements will provide benefit?

We therefore investigate six types of quality increases and use a range of talk-spurt lengths and transition times. In the next section we present a mathematical framework for the question. This framework drives the experiment design given in Section 3. Results and answers to the titular question are given in Section 4.

2. MATHEMATICAL BASIS

In order to address the question we begin with a mathematical basis. This basis must start with a statistical model for talk-spurt durations and we adopt the model provided in ITU-T Recommendation P.59 [10]. The talk-spurt lengths τ are specified to follow the exponential distribution with the probability density function

$$\begin{aligned} f(\tau) &= \lambda e^{-\lambda\tau},\ \ 0 \le \tau, \\ &= 0, \quad \text{otherwise.} \end{aligned} \tag{1}$$

The mean of this distribution is λ^{-1} and is defined to be 0.854 seconds in [10]. Measurements of real network traffic reported in [11] agree with this distribution and mean. The measurements of laboratory conversation tests found in [12] report higher talk-spurt means (between 1.4 and 1.6 seconds), and no distribution is reported.

Suppose a talk-spurt is in progress using Coder i which provides speech quality level Q_i. At time t, network resources give the opportunity to change to Coder j with speech quality level $Q_j > Q_i$. Should we switch to Coder j, or stay with Coder i? In a real-time decision-making environment we cannot know how long the current talk-spurt will last, but we could exploit the statistical model of talk-spurt durations given in (1). Specifically, we could calculate the expected talk-spurt quality for a talk-spurt starting with Coder i and switching to Coder j at time t:

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

$$\begin{aligned}\bar{Q}_{Ci,Cj}(t) =& E(Q_{Ci,Cj}(t,\tau))\\ =& \frac{\int_t^\infty Q_{Ci,Cj}(t,\tau)\lambda e^{-\lambda\tau}\,\mathrm{d}\tau}{\int_t^\infty \lambda e^{-\lambda\tau}\,\mathrm{d}\tau}.\end{aligned} \quad (2)$$

Given this result we could then compare with Q_i (and a possible cost function C) and switch only when an improvement is expected. That is, switch only when

$$\bar{Q}_{Ci,Cj}(t) \; > \; Q_i + C \, . \quad (3)$$

This approach requires Q_i, the static speech quality associated with Coder i, and $Q_{Ci,Cj}(t,\tau)$, the function that describes the speech quality of a length τ talk-spurt using Coder i for the first t seconds and Coder j for the remainder. Thus we are motivated to design, conduct, and analyze a subjective speech quality experiment that generates Q_i and $Q_{Ci,Cj}(t,\tau)$ for representative coder configurations Ci and Cj, transition times t, and talk-spurt durations τ.

The optional cost function C allows one to switch only when the expected quality increase exceeds any costs associated with making the switch from Coder i to Coder j. Note that these are costs that are not related to speech quality, but they must be expressed in speech quality units. This cost function is application-specific and is well outside the scope of this paper.

3. EXPERIMENT DESIGN

We designed a subjective speech quality experiment to measure the values of Q_i and $Q_{Ci,Cj}(t,\tau)$ required in (2). We included six types of coding transitions associated with five different speech coding modes. The NB and WB Adaptive Multi-Rate[1] (AMR) speech coders [13], [14] are prominent in wireless voice services. The SILK™speech coder [15] is used in the very popular Skype™ VoIP service and it offers medium band (MB) speech coding as well as NB and WB. Table 1 gives the details of the eight speech coding configurations used in this experiment. Note that the exact definitions of NB and WB vary by coder.

Coder	Passband	Rate (kb/s)	Rate (b/smp)
AMR NB Mode 0	85-3400 Hz	4.75	0.6
AMR NB Mode 2	85-3400 Hz	5.90	0.7
AMR NB Mode 7	85-3400 Hz	12.2	1.5
AMR WB Mode 1	50-7000 Hz	8.85	0.6
AMR WB Mode 8	50-7000 Hz	24.0	1.5
SILK NB	150-4000 Hz	8	1.0
SILK MB	80-6000 Hz	12	1.0
SILK WB	80-8000 Hz	16	1.0

Table 1. Speech coding configurations used. Speech passbands for AMR are nominal and [16], [17] show that the upper limit can change slightly with mode, at least when measured with tones or noise. SILK passbands are taken from Figure 4 of [3] and were measured with speech signals.

Table 2 defines the six conditions with speech coding transitions used in this experiment. Talk-spurts were coded by the entry coder (coder Ci with speech quality Q_i) for the first t seconds of the talk spurt, and by the exit coder (coder Cj with speech quality $Q_j, Q_i < Q_j$) for the final $\tau - t$ seconds. Each transition involves shifting to a higher quality speech coding mode (based on earlier work and confirmed here). We selected AMR modes for Conditions 1, 2, and 3 after consulting [2] where various AMR-NB and WB coding modes were evaluated on a 9-point absolute quality rating (ACR) scale. On that scale, each of the exit coders measured about 1.4 quality units above the corresponding entry coder. We selected SILK coding modes to consistently use one bit per sample. Note that Conditions 1 and 2 do not include increases in bandwidth, but the other 4 conditions do.

	Entry Coder (Ci)	Exit Coder (Cj)
1	AMR NB Mode 0	AMR NB Mode 7
2	AMR WB Mode 1	AMR WB Mode 8
3	AMR NB Mode 2	AMR WB Mode 1
4	SILK NB	SILK MB
5	SILK MB	SILK WB
6	SILK NB	SILK WB

Table 2. Transition conditions. Entry coder is used for first t seconds of talk-spurt and exit coder is used for the remaining $\tau - t$ seconds of talk-spurt.

We selected a five-point ACR experiment design, using the popular mean opinion score (MOS) speech quality scale. The speech material was designed to simulate talk-spurts from telephone conversations in North-American English. Toward that end we located transcripts of actual telephone conversations and excerpted suitable portions to form a list of talk-spurts. Two females and two males then recorded these talk-spurts in a sound-isolated chamber using studio-quality recording equipment. The model in (1) indicates that 99.9% of talk-spurts have a length of 6 seconds or less. Thus we located talk-spurts with approximate lengths of τ =1, 2, 3, 4, 5, and 6 seconds, resulting in six selected talk-spurts at each of these nominal lengths. These 36 selected talk-spurts were evenly divided between the female and male talkers (18 talk-spurts each) and approximately evenly divided between the individual talkers of each gender.

Next these talk-spurts were normalized to an active speech level of 26 dB below overload using [18], and then processed through the eight distinct speech coders and decoders listed in Table 1. The resulting speech files were then combined to create the six transition conditions shown in Table 2. We compiled, verified and ran version 1.0.8 (floating point) of Skype's SILK codec [15], [19], version 10.0.0 of the AMR NB and AMR WB codecs [13], [14], [20], [21], the filter tool in ITU-T P.191 [18], and version v14.3.2 of the SoX program [22] on a Mac Pro® 4,1 running Mac OS® X version 10.7. These tools, controlled by a set of Python™ (version 2.7.1) scripts created all processed talk-spurts.

We selected transition times of t = 0.5, 1, 2, 3, 4, and 5 seconds, measured from the start of the talk-spurt. A 2 ms crossfade was used at each transition to prevent waveform discontinuities and any associated auditory artifacts. All possible transition times t were paired with each talk-spurt duration τ, under the necessary constraint $t < \tau$ (21 combinations). Each τ value was repeated six times, once for each talk-spurt of that length. The experiment also included the eight coding conditions shown in Table 1 with no transitions. These static speech quality results are required because they provide the

[1] Certain commercial equipment, software, and services are identified in this report to specify adequately the technical aspects of the reported results. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the material or equipment identified is necessarily the best available for this purpose.

appropriate context for evaluating the transition conditions as seen in Section 4.

The experiment followed protocols set out in ITU-T Recommendation P.800 [23]. Speech recordings were converted to analog using a Benchmark® DAC1 digital to analog converter with a specified flat response (± 0.1 dB) from 20 Hz to 20 kHz. The listener could adjust the presentation level to the preferred level at any time, using a hardware knob on the converter. The presentation was diotic using Sennheiser HD 600 circumaural headphones with a specified -3 dB bandwidth from 16 Hz to 30 kHz. The experiment sessions were conducted in a sound-isolated room with background noise measured below 20 dBA SPL.

After each presentation the listener selected his or her opinion from one of five options: ("Excellent," "Good," "Fair," "Poor," and "Bad") using a graphical user interface (GUI). This GUI was presented on the touch-screen of an iPod touch®. The printed instruction at the top of the GUI was: "Please select your overall impression of the entire recording." When a selection was made, the experiment software proceeded with the next presentation. No replays were allowed.

Thirty listeners were recruited through random selection from the 1500 names listed in our research campus directory. Estimated ages ranged from 25 to 65, with a median of 35. Ten females and twenty males participated and none had knowledge of the experiment content. Each listener first completed a practice session with 16 presentations. This session exposed the listener to the full range of conditions, and allowed for verification of correct experiment operation. These scores were discarded. The main portion of the experiment was divided into two sessions with 111 trials in each. The median time spent on each session was 13 minutes. Every sixth listener heard the same files, but a different random presentation order was used for every listener. A total of 1332 (111 × 2 × 6) files were scored. 1188 of these were analyzed to develop the decision rules given in Section 4 and 144 files (11% of the total) were held in reserve and used only to test those decision rules.

4. RESULTS AND DISCUSSION

MOS values have a negative correlation with t, confirming that longer durations of entry quality lower the perception of overall quality. They have a positive correlation with $\tau - t$, indicating that longer portions of exit quality can compensate for more of the transition penalty. Example results for the case $\tau = 6$ seconds are shown in Figures 1 to 3. Each of these figures shows two of the transition conditions listed in Table 2 along with the associated two entry and exit conditions for reference.

The standard deviation of scores was very close to 0.96 for each of the conditions we tested. Each transition condition received 30 scores, resulting in 95% confidence intervals of ± 0.34 $(1.96\sigma/\sqrt{30})$. Static conditions received 180 or 360 scores (some served as both entry and exit conditions), resulting in 95% confidence intervals of ± 0.14 and ± 0.10. It is not our goal to make statements about significant differences in speech quality; rather we seek to analyze the underlying trends in mean speech quality as functions of t and τ. Thus, to preserve clarity, the figures show mean values only.

In each case we see that quality generally drops as the transition moves to a later time. Depending on the type of transition, late transitions can result in overall speech quality that is the same as or lower than the entry speech quality.

Next we reduce the large bank of results to simple decision rules as described in Section 2. The data collected provide empirical samples of Q_i and $Q_{Ci,Cj}(t, \tau)$ for $\tau = 1, 2, 3, 4, 5$, and 6, and $t = 0.5$,

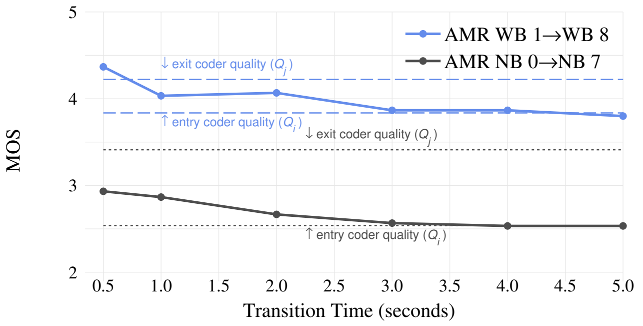

Fig. 1. Overall speech quality for 6 second talk-spurts with quality transitions: AMR WB 1 to 8 and AMR NB 0 to 7 (transition conditions 2 and 1 respectively).

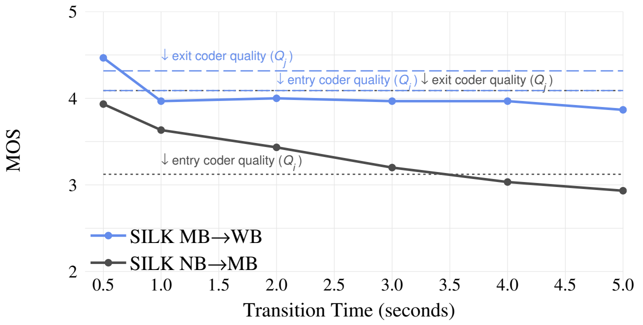

Fig. 2. Overall speech quality for 6 second talk-spurts with quality transitions: SILK MB to WB and SILK NB to MB (transition conditions 5 and 4 respectively).

1, 2, 3, 4, and 5. These samples can be used to calculate a discrete sum approximation to the integral in the numerator of (2):

$$\bar{Q}_{Ci,Cj}(t) \approx \widetilde{Q}_{Ci,Cj}(t) = \frac{\sum_{k=1}^{N} Q_{Ci,Cj}(t,\tau_k)p_k}{\int_t^{\infty} \lambda e^{-\lambda\tau}\,\mathrm{d}\tau}\ , \qquad (4)$$

$$\text{where } p_k = \int_{L_{k-1}}^{L_k} \lambda e^{-\lambda\tau}\,\mathrm{d}\tau\ ,$$

$$L_0 = t\ ,$$

$$L_k = \frac{\tau_k + \tau_{k+1}}{2},\ k = 1 \text{ to } N-1,$$

$$L_N = \infty\ .$$

The calculation of $\widetilde{Q}_{Ci,Cj}(t)$ given in (4) requires τ_k, $k = 1$ to N. These N values of τ are selected, in order, from the set of available τ values, $\tau \in \{1, 2, 3, 4, 5, 6\}$ to maximize N under the constraint that $t < \tau_k$. For example, if $t = 4$, then $N = 2$, $\tau_1 = 5$, and $\tau_2 = 6$.

We have calculated $\widetilde{Q}_{Ci,Cj}(t)$ for the six conditions presented in Table 2. In order to allow comparisons between these six results, we next normalized $\widetilde{Q}_{Ci,Cj}(t)$ to produce $\widehat{Q}_{Ci,Cj}(t)$:

$$\widehat{Q}_{Ci,Cj}(t) = \frac{\widetilde{Q}_{Ci,Cj}(t) - Q_i}{Q_j - Q_i}\ . \qquad (5)$$

Thus $\widehat{Q}_{Ci,Cj}(t)$ produces results on a normalized quality scale, where 0 corresponds to the entry quality (Q_i) and 1 corresponds to

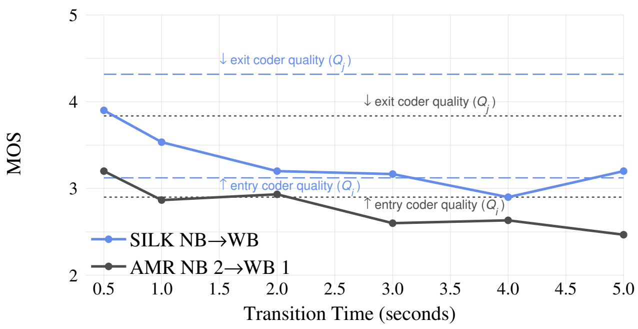

Fig. 3. Overall speech quality for 6 second talk-spurts with quality transitions: SILK NB to MB and AMR NB to WB (transition conditions 6 and 3 respectively).

the exit quality (Q_j). For the transition from SILK MB to SILK WB (Condition 5), the difference $Q_j - Q_i$ is very small (and not statistically significant in this experiment) and the division in (5) produces very noisy results. In essence, measuring speech quality values within this small interval requires greater measurement resolution than that attained in this experiment. We have excluded results from the SILK MB to SILK WB transition from the steps that lead to Figure 4.

The curves $\widehat{Q}_{Ci,Cj}(t)$ are similar for Conditions 1 and 2. These are the conditions where the quality increase does not include a bandwidth increase. Likewise they match closely for Conditions 3, 4, and 6—the conditions that include a bandwidth increase.

The result of averaging $\widehat{Q}_{Ci,Cj}(t)$ across Conditions 1 and 2 is shown by the dark line in Fig. 4. Averaging across Conditions 3, 4, and 6 gives the lighter line in that figure. The dashed line portrays the hypothetical case where overall speech quality is simply the duration-weighted average of the individual static speech qualities: $\tau^{-1}(tQ_i + (\tau - t)Q_j)$. This is included for reference and shows the perceived quality is indeed lower than this time-averaged quality, consistent with [9].

Each curve shows, as a function of t, the expected value of perceived talk-spurt quality when a transition is present t seconds into the talk-spurt. The expectation covers talk-spurts ranging from 1 to 6 seconds in length. The transition is from Ci to Cj and due to the normalization in (5), the perceived quality of Ci alone is 0, and the perceived quality of Cj alone is 1. These two curves summarize (across all talk-spurt lengths and conditions) the data gathered in this experiment in a form that provides answers to the titular question.

The dark curve in Fig. 4 remains above zero for $t < 2.8$ seconds. This means that on average, these quality increases (that do not include bandwidth increases) will increase the overall perceived quality only if they happen in the first 2.8 seconds of the talk-spurt. Otherwise they will slightly decrease the overall perceived quality. Note, however, that even switching coders a mere 0.5 seconds after the start of the talk-spurt garners only 56% of the speech-quality benefit associated with the second coder.

The lighter solid line in Fig. 4 remains above zero for $t < 1.8$ seconds. So, on average, these bandwidth-related quality increases will increase the overall perceived quality only if they happen in the first 1.8 seconds of the talk-spurt. Otherwise they will decrease the overall perceived quality. Switching coders 0.5 seconds after the start of the talk-spurt gives only 26% of the speech-quality benefit associated with the second coder. These results show that there is a penalty associated with these quality increases, and that penalty is larger when the quality increase includes a bandwidth increase.

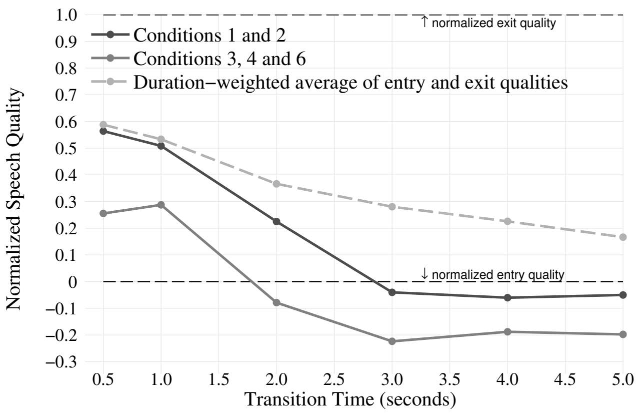

Fig. 4. Expected normalized speech quality for two classes of conditions and time averaging. Conditions 3, 4 and 6 include bandwidth increases and benefit is seen for increases made in first 1.8 seconds, Conditions 1 and 2 do not include bandwidth increases and benefit is seen for increases made in the first 2.8 seconds.

The thresholds 1.8 and 2.8 seconds give decision rules for speech quality increases. To account for a non-zero cost value C (see (3)) we can simply shift the curves in Fig. 4 downward by C. This will move the decision thresholds to earlier times and the resulting decision rules will guide us to speech quality increases that are "worth the cost."

For speech coding quality increases of the types and magnitudes used here, our answers to "When should a speech coding quality increase be allowed within a talk-spurt?" are as follows: If less than 1.8 seconds of talk-spurt have passed, these increases (either with or without bandwidth increases) are allowed and a benefit can be expected. If more than 1.8 seconds but less than 2.8 seconds of talk-spurt have passed, only the increases that do not include bandwidth increases are allowed and a benefit can be expected.

The subjective experiment included testing data that was was not used in the development of the results in Fig. 4 but was instead held in reserve for testing the decision rules developed. The testing data used the same four talkers as the development data, but used different talk-spurts, and the talk-spurt lengths ($1.25 \leq \tau \leq 5.5$) were selected to differ from the integer τ values used in the development data. Further, the transition times t were randomly selected and thus differed from the values used in the development process. The testing data covered the six conditions specified in Table 2 and was scored in the same subjective testing sessions as the development data. The testing data includes 96 different transition situations and represents about 11% of the subjective testing trials.

For the testing data the two decision rules derived above gave the correct answer (resulting in higher speech quality) in 70% of the cases (no bandwidth increase) and 68% of the cases (bandwidth increase). This performance is limited by the fact that the talk-spurt length τ cannot be known at decision time. For the same testing data the trivial rule "always switch to the higher quality coder" has somewhat lower performance and gave the correct answer in 59% and 55% of the cases respectively.

When the network conditions improve and a better speech coding option becomes available, then following the rules derived here gives greater certainty that the speech coding improvement will actually result in improved overall speech quality.

5. REFERENCES

[1] S. Voran, “Listener ratings of speech passbands,” in *Proceedings of the 1997 IEEE Workshop on Speech Coding for Telecommunications*, Sep. 1997, pp. 81–82.

[2] A. Ramo, “Voice quality evaluation of various codecs,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Mar. 2010, pp. 4662 –4665.

[3] A. Ramo and H. Toukomaa, “Voice quality evaluation of recent open source codecs,” in *Proc. Interspeech 2010*, Sep. 2010, pp. 2390–2393.

[4] S. Moller, M. Waltermann, B. Lewcio, N. Kirschnick, and P. Vidales, “Speech quality while roaming in next generation networks,” in *IEEE International Conference on Communications, ICC ’09*, Jun. 2009, pp. 1–5.

[5] B. Lewcio, M. Waltermann, S. Moller, and P. Vidales, “E-model supported switching between narrowband and wideband speech quality,” in *Proc. of the First International Workshop on Quality of Multimedia Experience, QoMEX 2009*, Jul. 2009, pp. 98–103.

[6] S. Voran, “Subjective ratings of instantaneous and gradual transitions from narrowband to wideband active speech,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Mar. 2010, pp. 4674 –4677.

[7] L. Gros and N. Chateau, “Instantaneous and overall judgements for time-varying speech quality: Assessments and relationships,” *Acta Acustica united with Acustica*, vol. 87, pp. 367–377, 2001.

[8] P. Gray, R. Massara, and M. Hollier, “An experimental investigation of the accumulation of perceived error in time-varying speech distortions,” in *Preprint, Audio Engineering Society 103rd Convention*, New York, 1997.

[9] S. Voran, “A basic experiment on time-varying speech quality,” in *Proc. of the 4th International MESAQIN (Measurement of Speech and Audio Quality in Networks) Conference*, Prague, Czech Republic, Jun. 2005, pp. 51–64.

[10] *Artificial Conversational Speech*, ITU-T Recommendation P.59, 1993.

[11] P. Pragtong, T. Erke, and K. Ahmed, “Analysis and modeling of VoIP conversation traffic in the real network,” in *Fifth International Conference on Information, Communications and Signal Processing*, 2005, pp. 388 –392.

[12] F. Hammer, P. Reichl, and A. Raake, “Elements of interactivity in telephone conversations,” in *Proc. Interspeech 2004*, 2004.

[13] *Mandatory speech CODEC speech processing functions; AMR speech Codec; General description*, ETSI/3GPP TS 26.071, Rev. 10.0.0, Apr. 2011.

[14] *Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description*, ETSI/3GPP TS 26.171, Rev. 10.0.0, Apr. 2011.

[15] K. Vos, S. Jensen, and K. Soerensen, “SILK speech codec,” RFC draft-vos-silk-02, IETF, Sep. 2010. [Online]. Available: http://tools.ietf.org/html/draft-vos-silk-02

[16] ETSI, “Performance characterization of the adaptive multi-rate (AMR) speech codec,” ETSI, Tech. Rep. TR 126 975, Jan. 2009.

[17] ——, “Performance characterization of the adaptive multi-rate wideband (AMR-WB) speech codec,” ETSI, Tech. Rep. TR 126 976, Jan. 2009.

[18] *Software tools for speech and audio coding standardization*, ITU-T Recommendation P.191, 2005.

[19] (2011) SILK. Website. Skype. [Online]. Available: http://developer.skype.com/silk

[20] *ANSI-C code for the floating-point Adaptive Multi-Rate (AMR) speech codec*, ETSI/3GPP TS 26.104, Rev. 10.0.0, Apr. 2011.

[21] *Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; ANSI-C code*, ETSI/3GPP TS 26.204, Rev. 10.0.0, Apr. 2011.

[22] (2011) SoX - Sound eXchange. Website. [Online]. Available: http://sox.sourceforge.net/

[23] *Methods for subjective determination of transmission quality*, ITU-T Recommendation P.800, 1996.