The Well-Tempered Conversation: Interactivity, Delay and Perceptual VoIP Quality

Florian Hammer and Peter Reichl
Telecommunications Research Center Vienna (ftw.)
Vienna, Austria
Email: {hammer|reichl}@ftw.at

Alexander Raake
Institute of Communication Acoustics (IKA),
Ruhr-University Bochum, Germany
Email: alexander.raake@ruhr-uni-bochum.de

Abstract— **The factors causing perceptual quality impairment on Voice-over-IP (VoIP) connections include traditional network Quality-of-Service (QoS) parameters like packet loss rate, delay or jitter as well as parameters characterizing the conversation itself. Among the latter ones, we focus on the impact of "conversational interactivity" on the perceptual quality of a phone conversation. We introduce "Parametric Conversation Analysis" as a formal framework for the instrumental investigation of conversational parameters at different transmission delay conditions, we further present the notion of "conversational temperature" as an intuitive scalar metric for the interactivity of conversations, and we demonstrate the application of our methods to a set of conversation recordings performed under various delay conditions, also with respect to results of subjective quality ratings.**

Keywords–**VoIP, Perceived QoS, Delay, Parametric Conversation Analysis, Conversational Interactivity, Conversational Temperature**

I. Introduction and Related Work

Today's communication technology is characterized by the shift from circuit-switched to packet-switched networks merging data transport with real-time applications. One of the hottest topics in this area is Voice-over-IP (VoIP) which is, in the long run, expected to replace the traditional landline telephone. However, the convergence of data networks and telephone networks introduces certain degradations with regard to the end-to-end speech quality perceived by the user. During data transmission bursts like extensive download activities, voice packets are likely to be lost, which introduces an audible quality impairment. Moreover, the entire chain of signal processing and networking introduces a considerable amount of end-to-end delay. Depending on the interactivity of the conversation, such a transmission lag may disturb the natural flow of the dialogue and thus degrade the perceived Quality-of-Service (QoS). Summarizing, we are interested in the conversational factors which affect the user's quality perception of a VoIP call.

The novel contributions of this paper are threefold: we introduce a framework for a parametric analysis of telephone conversations, we present a metric for determining conversational interactivity in terms of a "conversational temperature", and we finally apply this methodology to the analysis of recorded telephone conversations performed under different delay conditions. As a major result we analyze the correlation of the conversational parameters with the quality ratings given by the test persons.

The measurement of the perceived degradation caused by delay requires subjective conversational speech quality tests. Kitawaki [1], and more recently, Möller [2], Takahashi [3] and Raake [4] have investigated the impact of delay on the perceived speech quality. A major methodological difference between these studies is the type of conversation tasks that have been used for the tests. As an example, the rapid exchange of random numbers as used by Kitawaki results in a completely different conversational structure and interactivity than tasks like reserving a plane ticket as used by Möller and Raake. Therefore, we aim for identifying the conversational parameters which most dominantly influence the users' quality ratings and for finding a metric which supports us to distinguish the levels of interactivity of test conversations.

It is important to note that delay is not audible - in terms of a "listening-only" degradation - if there is no echo on the line or the echo is successfully suppressed by an echo canceller. In fact, the end-to-end delay impairment is, if at all, perceived within the course of a conversation which results in delayed responses and the potential inability to interrupt the conversation partner. Issues regarding the conversation structure are usually addressed in Conversation Analysis (CA) by studying the turn-taking process between the conversation participants [5]. Ruhleder and Jordan have investigated the influence of delay on the distributed (video-mediated) interaction [6]. They conclude that people behave in ways that violate the expectations of their conversational partner rather than breaking conversational rules [6] when it comes to transmission lags.

We approach this matter by extending Brady's concept of conversational events like "talk spurt", "double talk", and "interruption" [7]. The set of conversational parameters we utilize can be extracted from conversation recordings in an instrumental way. This procedure allows a detailed analysis of the relation between transmission delay, interaction parameters and the quality as perceived by the user.

The remainder of this paper is structured as follows: Section

II introduces the concept of *Parametric Conversation Analysis* which we use to analyze recorded telephone conversations in an instrumental way. Section III deals with *"Conversational Temperature"* as a simple and intuitive metric describing the interactivity of a conversation. Section IV provides an overview on measurement and evaluation aspects, before Section V presents selected central results in detail. Finally, section VI concludes the paper with a summary and an outlook on future work.

II. Parametric Conversation Analysis

In this section, we introduce the concept of *Parametric Conversation Analysis* (PCA). The term "parametric" reflects the fact that this framework is concerned with conversational parameters that can automatically be extracted from recorded conversations. We may further distinguish between PCA-T, PCA-F and PCA-M, as far as telephone, face-to-face or multimedia conversations are concerned, respectively. In [8], we have already used the PCA-T approach to compare two types of scenarios used in conversational speech quality assessment. In the remainder of this paper, we focus on PCA-T and analyze the parameters of a 4-state conversation model and a corresponding set of conversational events in Section II-A, whereas Section II-B extends this approach by integrating the impact of a transmission channel showing delay. Note that in the course of the paper, parameters from both the conversational model and the conversational events together are referred to as "conversational parameters".

A. Conversational Model and Conversational Events

Following [9], we model the structure of a two-way conversation by the distinction of four different states as illustrated in Fig. 1. States A and B denote that either person A or person B is talking and the other person is silent. State M (mutual silence) refers to both persons being silent, and state D (double talk) represents the case of both persons talking simultaneously. Heading a step further and including the transitions between these four states, we end up with a stochastic process as depicted in Fig. 2. Note that direct transitions between A and B as well as between M and D represent extremely rare events (normally each talk spurt is either followed by mutual silence or interrupted by a period of double-talk) and may therefore be omitted.

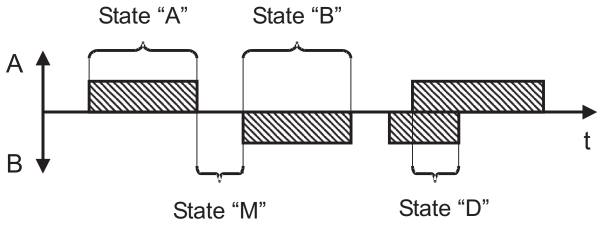

Fig. 1. Conversational States.

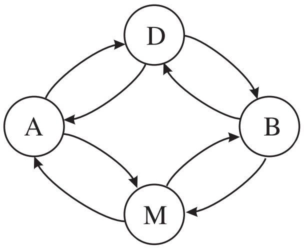

Fig. 2. General Conversation Model.

Distinguishing the different "conversational events" in analogy to [7] allows to describe the characteristics of a conversation in more detail. To this end, we define a *speaker alternation* to be the switching event between the two speakers which is separated by mutual silence (i.e. state sequences *A-M-B* and *B-M-A*), whereas an *interruption* corresponds to the first speaker being interrupted by the second one who eventually (i.e. after a period of double-talk) gains the floor and continues to talk alone (i.e. state sequences *A-D-B* and *B-D-A*). The *Speaker Alternation Rate* (SAR) represents the total number of speaker alternations plus interruptions per minute, and similarly the *Interruption Rate* (IR) is defined as the number of interruptions per minute. Further relevant conversational events include the *pause* as mutual silence period between two talkspurts of the same speaker (i.e. state sequences *A-M-A* or *B-M-B*, resp.), and *non-interruptive double-talk* (i.e. state sequences *A-D-A* or *B-D-B*) in which the original speaker does not give up the floor despite of an intermediate double-talk period.

B. The Impact of Transmission Delay

The introduction of a delayed transmission channel complicates the situation described in the previous section significantly. The delayed transmission of immediate responses and (intended) interruptions shuffles the conversational structure and may irritate the participants. Response times that are increased by a large round-trip time may put the participants' patience to test, especially in situations which demand high interactivity. However, users need to adapt their behavior in order to be able to nonetheless have a "normal", disciplined conversation.

On a more formal level, it is required to distinguish three different state patterns, depending on whether we observe the states at speaker A, at speaker B or on an absolute time scale. Fig. 3 illustrates a simple example for the differences between the state sequence patterns with respect to A (above), to B (below) and to the absolute clock (depicted within the transmission channel). In this example, speaker B receives speaker A's delayed utterance and responds after a certain think time. Note that, in total, B's response as perceived by talker A is delayed by one round-trip time. After some time, A starts to talk assuming that B is not responding to her first talk-spurt. In effect, the delayed utterance of B interrupts talker A (without intention). At talker B's side, B is interrupted by A before herself interrupting A on purpose. Altogether, it is quite interesting to observe the different state sequences as perceived by speaker A and speaker B, and in contrast to the time stamps as registered by the absolute clock in-between (cf. [6]).

As a first consequence, we have to distinguish two different types of interruption events: an *active interruption* is an intended interruption or a natural overlap performed by one of the speakers at her own side (i.e. while she is still listening to the other speaker), and a *passive interruption* as the unintended event of being interrupted by the other speaker while speaking myself. Both cases are depicted in Fig. 3.

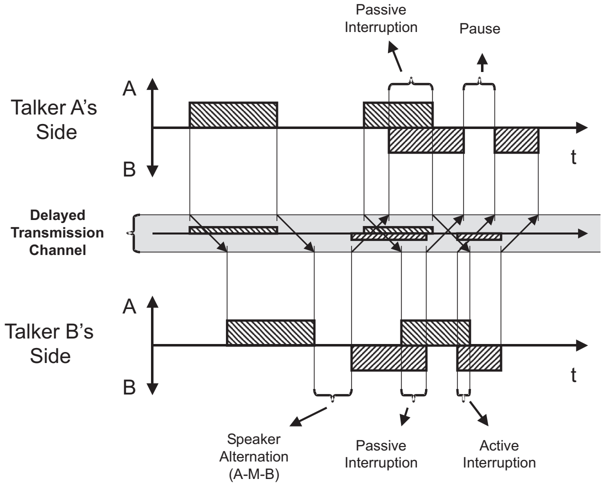

Fig. 3. Impact of Transmission Delay on the Conversational Structure.

The delayed responses lead to a total increase in mutual silence at both ends which is expected to be proportional to the delay time. Moreover, we expect an increased amount of passive interruptions at higher delays due to the shuffling caused by the delay. An analysis of the mean durations of speaker alternation and interruption and the amount of non-interruptive double-talk is not performed in this paper but is considered for future work.

III. A Temperature Metric for Conversational Interactivity

Coming back to the conversation model described in Section II-A, for each state $I \in \{A, B, M, D\}$, let t_I be the average sojourn times spent in these states, with $t^* =$ max $\{t_A, t_B, t_M, t_D\}$ being their maximum. In this section, we derive a scalar parameter $\tau = \tau(t_A, t_B, t_M, t_D)$ as function of these mean sojourn times, leading to a simple but efficient and intuitive one-dimensional metric for describing conversational interactivity.

As related work still has not managed to agree on a satisfying *explicit* definition of interactivity, in [10] we have proposed an *implicit* approach, introducing three descriptive axioms which characterize central features of conversational interactivity. In the present paper, we refine this idea and its implications by presenting a much more formal version. In this sense, the mentioned axioms read now as follows:

Axiom I (Limiting behavior):

$$\lim_{t^* \to \infty} \tau(t_A, t_B, t_M, t_D) = 0 \quad \textit{and} \tag{1}$$

$$\lim_{t^* \to 0} \tau(t_A, t_B, t_M, t_D) = \infty. \tag{2}$$

Axiom I suggests that the conversation is *not* interactive at all if either A or B are speaking all the time, no one is speaking at all, or both speakers are simultaneously active all the time. On the other hand, the case of high interactivity corresponds to state sojourn times being short.

Axiom II (Normalization):
Norm (average) conversations have average interactivity $\bar{\tau}$.

Axiom II scales our interactivity metric alongside an abstract "average conversation" with sojourn times averaged over many different conversation samples.

Axiom III (Monotonicity/First-order Behavior):

$$t_I((1 + f(\tau)\epsilon) \cdot \tau) = (1 - \epsilon) \cdot t_I(\tau) \tag{3}$$

for $I \in \{A, B, M, D\}, \epsilon \to 0$ and $f(\tau) > 0, \tau > 0$.

Finally, Axiom III implicates monotonicity of τ in the sense that decreasing sojourn time in one of the states leads to an increase of the interactivity metric and vice versa. Note that $f(\tau)$ is supposed to be a positive real function describing the local first-order behavior of the state sojourn times as a function of the interactivity of the conversation, and could, for instance, be chosen from the family $const \cdot \tau^\alpha$, $\alpha \in \mathbb{R}$. Axiom I poses certain marginal conditions on this function, e.g. $\alpha \geq 0$. As a first choice, $\alpha = 0$, i.e. $f(\tau) \equiv const$, leads straightforward to a solution $t_I(\tau) \propto \frac{1}{\tau}$. However, we will demonstrate in the rest of the section that the "linear" choice $f(\tau) = \tau/\bar{\tau}$ (i.e. $\alpha = 1$), linking together Axioms II and III, leads to an especially appealing form for the desired interactivity metric.

Remembering that we aim at an explicit expression for τ, we recognize that for our choice of $f(\tau) = \tau/\bar{\tau}$, the Taylor expansion of t_I in (3) leads to

$$\frac{d}{d\tau} t_I(\tau) = -(\frac{\bar{\tau}}{\tau^2}) \cdot t_I(\tau). \tag{4}$$

Solving this differential equation, we get

$$t_I(\tau) \propto exp(\frac{\bar{\tau}}{\tau}). \tag{5}$$

Now we interpret Fig. 2 as a regular Continuous Time Markov Chain (CTMC) and define $1/\nu_I$ to be the constant of proportionality in (5). Then, the sojourn time in state I of the CTMC is exponentially distributed [11] with parameter

$$\lambda_I = \frac{1}{t_I} = \nu_I \cdot exp(-\frac{\bar{\tau}}{\tau}) \tag{6}$$

λ_I can be interpreted as total transition rate out of state I. This leads us to a related class of problems well-known from statistical thermodynamics [12]: Imagine a single particle moving within a quantum well bordered by potential walls where the particle continuously tries to jump over one of the walls. An important parameter describing this system is its temperature T, and the success rate of the jumping particle depends on T according to

$$\lambda = \nu \cdot exp(-\frac{\Delta E}{kT}). \tag{7}$$

Here, ΔE describes the height of the potential walls, ν is the oscillation frequency of the particle, and k is known as "Boltzmann's constant". Comparing (6) and (7) suggests an interpretation of the interactivity metric τ in terms of a temperature, the so-called "conversational temperature" as proposed in [10].

From here, it is left to determine the parameter ν_I. From Axiom II and (6) we learn that a norm conversation, with sojourn time $\bar{t}_I$ in state I, leads to

$$\nu_I = e/\bar{t}_I \tag{8}$$

Summarizing (6) and (8) and solving for τ yields

$$\tau = \tau_I(t_I) = \frac{\bar{\tau}}{ln(t_I) - ln(\bar{t}_I) + 1}, I \in \{A, B, M, D\} \tag{9}$$

But we can use (9) also the other way round: Assume $\tilde{t}_I$ to be the measured average sojourn time in state I for an arbitrary conversation. In this case, (9) provides a direct way of estimating the temperature of the conversation by calculating $\hat{\tau}_I = \tau_I(\tilde{t}_I)$.

Applying this estimation procedure to all four states $I \in \{A, B, M, D\}$ in general may lead to four slightly different values for τ, due to the statistical nature of our conversation model. Therefore, we use a least-square fitting approach to finally end up with one uniform scalar parameter $\hat{\tau}$ estimating the conversational temperature of the conversation as such:

$$\hat{\tau} = argmin \sum_I (\bar{t}_I \cdot exp(\frac{\bar{\tau}}{\tau} - 1) - \tilde{t}_I)^2 \tag{10}$$

Finally, we have to quantify $\bar{\tau}$ from Axiom II. For the sake of simplicity, in the remainder of the paper we choose the conversational temperature of a norm conversation to be "*room temperature*", i.e. 21.5° (with respect to the Celsius scale and thus corresponding to 294.65 K or 70.5° F).

IV. Measurements and Evaluations

A. Measurement Setup

Our results as presented in Section V are mainly based on conversations that have been recorded during speech quality tests carried out at the Institute of Communication Acoustics (IKA) at Ruhr-University Bochum. The laboratory setup for these tests is described in [2] and [13]. The conversational tests consisted of VoIP connections using the ITU-T G.729 codec with different bursty packet loss rates (0%, 3% 5% and 15%) combined with transmission delay of 60 ms, 360 ms, 660 ms and 960 ms. In this study, we restrict ourselves to scenarios without packet losses in order to explore the pure delay effect on both interactivity and subjective quality. The quality ratings comprise both mean opinion scores (MOS) based on the 5-point absolute category rating scale recommended by the ITU-T [14], which reflect the overall quality perceived by the user and range from 5.0 ("excellent") to 1.0 ("bad"). The CR-10 category rating scale [15], [2] indicates the perceived impairment of the connection: here, a value of "0" denotes that the user has not perceived any impairment at all, whereas, e.g. "2", "5" and "10" correspond to "weak", "strong", and "extremely strong" impairment, respectively. Note, that the CR-10 scale is copyrighted by the author [15].

The test subjects were asked to accomplish interactive Short Conversation Tests (iSCT, [13]) which represent telephone scenarios like the rapid exchange of information about new employees, e.g. e-mail addresses and telephone numbers, leading to comparable and balanced conversations of higher interactivity compared to the standard Short Conversation Test (SCT) scenarios [2]. The iSCTs were designed to create a conversation situation in which the participants would be more sensible for increased transmission delay.

For our investigations, we have recorded the microphone signals of both speakers and manually coded the talkspurts in order to reach high accuracy in the derived parameters. The talk spurts were shifted in time as to obtain the conversational patterns as perceived by the individual participants.

B. Evaluation

For the evaluation we have re-ordered the speakers according to their average speech activity (i.e. ratio between total sum of talk-spurts and total duration of the conversation). Therefore, generally speaker A is the one talking more. We have evaluated both the conversational parameters resulting from the PCA-T and the conversational temperature and put them into relation with the subjective quality ratings. We mainly focus on the evaluation of the parameters for increasing transmission delay times.

V. Results

In the following, we present the analysis of conversations performed by 7 pairs of test persons (8 female, 6 male) who knew each other. We consider the average call duration of

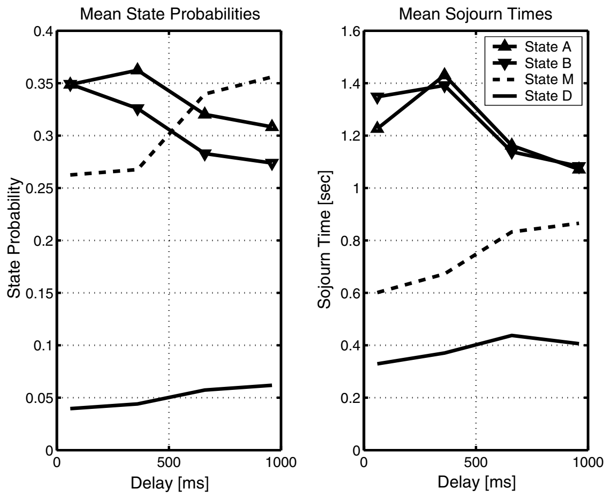

Fig. 4. Conversational Model Parameters versus Delay.

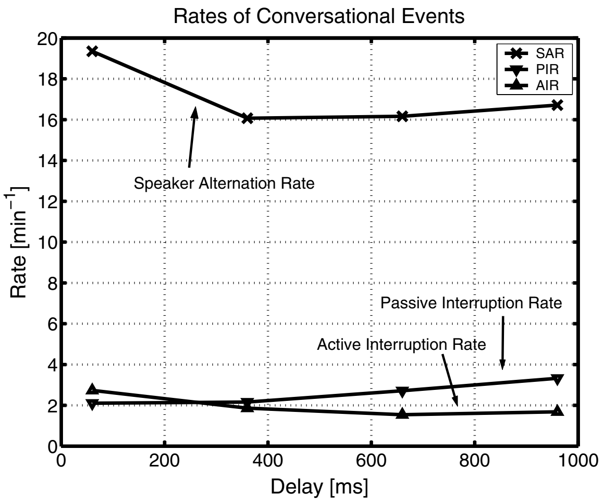

Fig. 5. Rates of Speaker Alternation (SAR), Active Interruption (AIR), and Passive Interruption (PIR).

143 sec as long enough for the test persons to obtain a useful impression of the properties of the connections.

A. Conversational Parameters

Our first analysis addresses the behavior of the conversational parameters at the different delay conditions. Fig. 4 presents the model parameters. As a first observation, the mean state probabilities and sojourn times of states *A* and *B* are decreasing with increasing delay, while the values for mutual silence (*M*) show a significant increase between 360 ms and 660 ms. The latter behavior is not unexpected, because we already suspected mutual silence to increase with delay due to the time lag of the responses (see Section II-B). However, the present data suggests a change in behavior of the subjects when the delay is increased from 360 ms to 660 ms. In contrast to the variation of these parameters, double-talk (D) rises only slightly with delay.

Fig. 5 illustrates the rates of the three conversational events introduced in Section II-B. The Speaker Alternation Rate (SAR) starts with 20 per minute at a delay of 60 ms and remains fairly stable around 16 min^{-1} for higher delay values. The Active Interruption Rate (AIR) decreases (thus, at least some of the participants seem to notice the lag and adapt their conversational behavior). As already expected in Section II-B, transmission delay increases the Passive Interruption Rate (PIR) due to the uncontrolled shuffling of the utterances caused by the delay.

B. Temperature

As already sketched in Fig. 3, the presence of delay causes significant differences concerning the on/off patterns for the left and right speaker of a conversation. Therefore, the temperature metric introduced in Section III needs to be checked for consistency first, as the interactivity of an individual conversation must be independent of the speakers' perspectives. Fig. 6 depicts the temperature for delays of 360 ms and 960 ms, resp., from the perspectives of both speakers and for seven different conversations each. The results demonstrate that for low delay (360 ms), the temperature derived from the A's and B's on/off patterns (labelled A_360 and B_360) are roughly identical for most conversations, whereas for higher delay (960 ms), the curves do no longer coincide, but the consistency between both temperatures is still preserved. Therefore, we may consider the temperature of any given conversation as uniquely determined.

Furthermore, Fig. 6 shows also that for roughly half of the conversations, the temperature for the case of 960ms delay is significantly higher than for the case of 360ms. This is true also for the average conversational temperature as is demonstrated in Fig. 7 for all four investigated delays. Here, we observe that the temperature minimum of 19.9° is attained for a delay of 360 ms, whereas the subsequent rise towards 23.7° at 960 ms is significant also w.r.t. the depicted standard deviations.

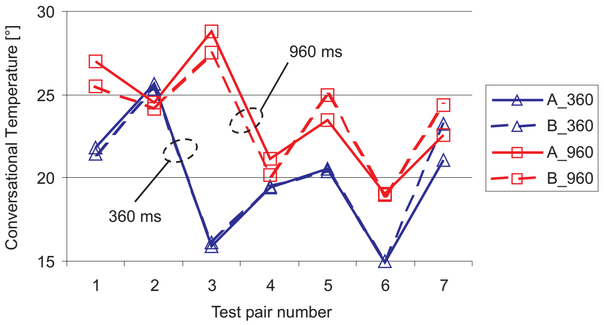

Fig. 6. Consistency of the Temperature Metric between Speakers A and B for Delays of 360ms and 960ms.

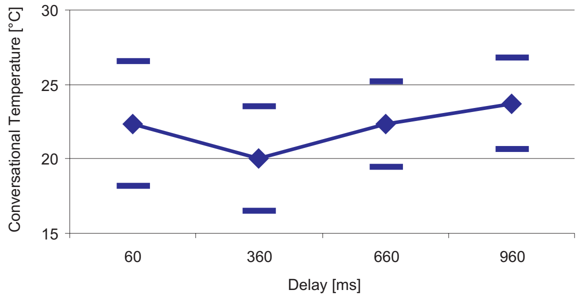

Fig. 7. Average Temperature vs. Delay and Standard Deviations.

C. Perceptual Quality

Regarding the correlation of the speech quality perceived by the test persons with both the conversational parameters and the conversational temperature at different delay times, Fig. 8 compares the evolution of MOS and CR-10 value, Passive Interruption Rate and Conversational Temperature with respect to increasing delay, where each parameter has been averaged over all conversations. Both the MOS and CR-10 ratings indicate only a slight decrease in perceptual quality at very high delay. Note as an important observation that the PIR are strongly correlated with both MOS and CR-10 scales for larger delay values. Moreover, this increase of the PIR may lead to the significant increase in conversational temperature, given the fact that the SAR remains fairly stable.

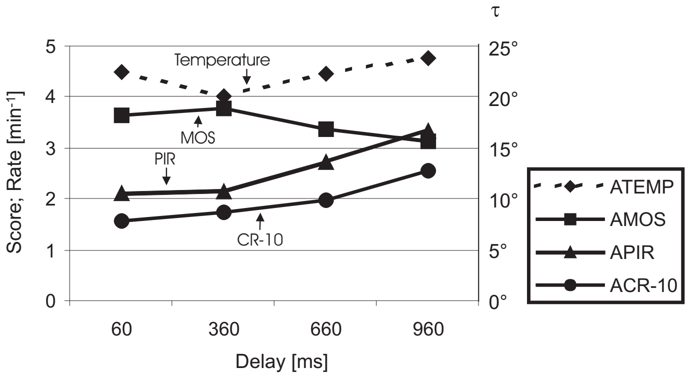

Fig. 8. Comparison of Average Temperature (ATEMP), Average Passive Interruption Rate (APIR), Average MOS (AMOS) and Average CR-10 vs. Delay.

VI. Conclusions

This paper has introduced two approaches for formally describing telephone conversations, i.e. the framework of parametric conversation analysis and a temperature metric for conversational interactivity. To this end, we have analyzed a total of 28 conversations carried out over connections with different amounts of delay.

Our results show that high delay has only a slight impact on the quality perceived by the users. The analysis of the conversational parameters indicates a high correlation between the quality rating and the occurrence of passive interruptions. Both the increase in conversational temperature and the stable value of the speaker alternation rate in combination with the increase of the passive interruption rate explain the fact that the interactivity of the conversations in terms of the conversational temperature increases with delay.

Future work focusses on the integration of both mentioned approaches into one instrumental interactivity metric which eventually allows the distinction of conversational tasks used for conversational speech quality assessment. Complementary, we plan to develop a measurement methodology for "perceived interactivity". Thus, we will be able to obtain a more comprehensive picture about telephone users, their conversational habits and their perception of transmission delay.

Acknowledgments

This work has been partially funded under the Austrian government's Kplus Competence Center Program, and has partially been funded in the framework of the EU-IST project INSPIRE (IST-2001-32746, http://www.inspire-project.org).

References

[1] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications," *IEEE J. Sel. Areas Comm.*, vol. 9, no. 4, pp. 586–593, May 1991.
[2] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications*. Boston: Kluwer Academic Publishers, 2000.
[3] A. Takahashi, "Opinion model for estimating conversational quality of VoIP," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 3, Montral, Canada, May 2004, pp. 1072–1075.
[4] A. Raake, "Predicting speech quality under random packet loss: Individual impairment and additivity with other network impairments," *ACUSTICA/Acta Acustica*, vol. 90, no. 6, pp. 1061–1083, Nov/Dec 2004.
[5] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.
[6] K. Ruhleder and B. Jordan, "Meaning-making across remote sites: How delays in transmission affect interaction," in *Proceedings of the Sixth European Conference on Computer-Supported Cooperative Work*, Cophenhagen, Denmark, 1999, pp. 411–427.
[7] P. T. Brady, "A statistical analysis of on-off patterns in 16 conversations," *Bell System Technical Journal*, vol. 47, no. 1, pp. 73–91, Jan. 1968.
[8] F. Hammer, P. Reichl, and A. Raake, "Elements of interactivity in telephone conversations," in *8th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004)*, Jeju Island, Korea, Oct. 2004.
[9] International Telecommunication Union, "Artificial conversational speech," *ITU-T Recommendation P.59*, Mar. 1993.
[10] P. Reichl and F. Hammer, "Hot discussion or frosty dialogue? Towards a temperature metric for conversational interactivity," in *8th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004)*, Jeju Island, Korea, Oct. 2004.
[11] S. M. Ross, *Stochastic Processes*. Wiley, 1996.
[12] K. Stowe, *Introduction to Statistical Mechanics and Thermodynamics*. Wiley, 1983.
[13] International Telecommunication Union, "E-model: Additivity of burst loss impairment with other impairment types," *Source: Ruhr-University Bochum (A. Raake), ITU-T Delayed Contribution 221*, 2004.
[14] ——, "Methods for subjective determination of transmission quality," *ITU-T Recommendation P.800*, Aug. 1996.
[15] G. Borg, *Borg's Perceived Exertion and Pain Scales*. Champaign, IL: Human Kinetics, 1998.