A General Temperature Metric Framework for Conversational Interactivity

Peter Reichl[1], Gernot Kubin[2], Florian Hammer[1]

[1]Telecommunications Research Center Vienna (ftw.), Donaucitystr. 1, A-1220 Vienna, Austria
[2]Signal Processing and Speech Communication Laboratory, TU Graz, Inffeldgasse 16c, A-8010 Graz, Austria
{reichl | hammer}@ftw.at; gernot.kubin@tugraz.at

Abstract. *In order to investigate the potential impact of interactivity on the perceived quality of future telecommuncation services, it is necessary to define a simple but efficient metric for this conversation parameter. To this end, recently it has been proposed to use an obvious analogy to statistical thermodynamics and define interactivity as system temperature of a Markov chain describing the conversation. The present paper refines and extends this approach towards a general framework for the resulting temperature metric. We describe several state models for conversations providing for interactivity-specific requirements. Based on that, an additional parameter is assigned to each state in order to describe its specific impact on the interactivity metric properly. We derive an explicit solution of the resulting metric, before we finally discuss how to deal with the case of more than two participants by introducing a new entropy model for conversational temperature.*

Keywords: Conversational interactivity, entropy, perceptual speech quality, E-model

1 Introduction

The question of how to integrate the concept of "conversational interactivity" into instrumental metrics for perceptual Quality-of-Service (QoS) like ITU-T's E-model [6] recently has raised considerable interest both within standardization bodies and the research community. There is, however, a fundamental prerequisite for describing the impact of interactivity onto the service quality as perceived by the end user: we need to measure interactivity in a technically simple, but nevertheless expressive way. Related work has repeatedly tried to describe this central conversation parameter through various parameters, ranging from conversational events [4] over prosodic elements (e.g. changing voice frequency) [1] towards semantic approaches like [8]. Most of these attempts, though, presuppose the existence of a generic unified definition of the concept of interactivity itself, which, unfortunately, is still not available in the related work [2].

Therefore we have argued in [9] that, if we cannot define explicitly what interactivity is, we should start by describing *what it is not*. The basic assumptions of this indirect approach are reviewed and slightly refined in Section 2, before we describe the resulting scalar interactivity metric which is closely related to the well-known physical notion of temperature in a thermodynamical system.

Whereas in [3] we have derived such a temperature metric as an overall system parameter τ using an implicit non-linear optimization procedure, the present paper assigns temperatures τ_I, sojourn probabilities π_I and a specific weight c_I to each single conversation state I, and describes the resulting overall metric τ as weighted average (with normalization constant N):

$$\tau = \frac{1}{N} \sum c_I \cdot \pi_I \cdot \tau_I \ . \qquad (1)$$

Beyond this general change of perspective, the main contributions of this paper focus on extending the original approach towards a much more general framework for measuring interactivity. Section 3 deals with the question of how to describe a interactive conversation in terms of conversation states and discusses some examples of resulting models of different complexity. In Section 4, we introduce a new state-specific parameter which characterizes the impact of each individual state on the perceived interactivity, and demonstrate that this parameter is closely related to the thermodynamic notion of specific heat capacity. Section 5 summarizes the resulting explicit metric, thus simplifying the implicit description of [3] significantly. Finally, in Section 6 we sketch an entropy-related approach to further extend the metric towards the case of more than two persons participating in a conversation. Section 7 concludes the paper with a couple of summarizing remarks.

2 Conversational Temperature as an Interactivity Metric

In our search for a scalar metric τ which allows to describe the interactivity of a conversation in a simple, efficient but yet intuitive way without requiring extensive technical measurement equipment, we follow [9] and [3] in refraining from an explicit definition of interactivity and a respective metric. Instead, we use an indirect approach in close analogy to related models in statistical thermodynamics and characterize τ as the "temperature" of the conversation. To this end, we assume that the conversation can be modeled as a sequence of states I taken from a set σ and transitions ϑ_{IJ} between them, $I, J \in \sigma$, and focus on the average sojourn times t_I spent in these states as the only input parameters for τ:

$$\tau = \tau(t_I)_{I \in \sigma} . \qquad (2)$$

Figure 1 depicts a standard conversation model (Model 1) taken from [5] which consists of the four states "*A*" (speaker A active, speaker B silent), "*B*" (A silent, B active), "*D*" (double talk, i.e. A and B active at the same time) and "*M*" (mutual silence, i.e. neither A nor B active).

Now we define $t = (t_I)_{I \in \sigma}$ to be the vector of average sojourn times spent in the various states I, and call their maxi-

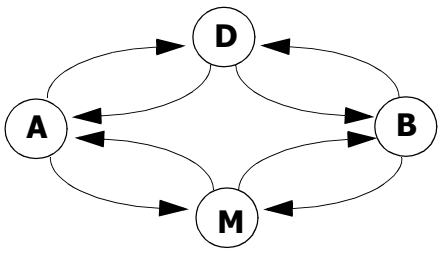

Figure 1: Conversation Model 1 (4 States)

mum $t^* = max\{t_I | I \in \sigma\}$. Using the standard (game-theoretic) notation of t_{-I} for the vector consisting of all t_J except for t_I, i.e. $J \in \sigma \backslash \{I\}$, $t_I(\tau; t_{-I})$ describes the dependency of t_I on the interactivity τ if all other t_J, $J \neq I$, are fixed.

Note that the mentioned indirect approach basically boils down to an appropriate description of the *lack of interactivity* in terms of the individual t_I's. In [9], we have already argued that in Model 1 there is reduced interactivity if (a) speaker A (or speaker B) is continuously holding monologues, (b) if both speakers are active simultaneously and thus have difficulties to react to the information flow between them, and (c) if both speakers are silent and there is no information flow at all. We will see later in Section 4 that these three cases affect interactivity to different degrees, but nevertheless we may formally describe the limiting case of this assumption as follows:

$$\lim_{t_I \to \infty} \tau(t_I; t_{-I}) = 0 \text{ for } I \in \sigma, \tag{3}$$

which implies the version originally proposed in [3]:

$$\lim_{t^* \to \infty} \tau(t_I; t_{-I}) = 0. \tag{4}$$

On the other hand, the interactivity increases if *all* t_I tend to be rather short, i.e. (in the limiting case):

$$\lim_{t^* \to 0} \tau(t_I; t_{-I}) = \infty. \tag{5}$$

For the full description of our framework, (3) and (5) have to be complemented by two further assumptions: first of all, we need to scale the resulting temperature metric, and the easiest approach is to take a certain "norm conversation" (with state sojourn times averaged over a large set of typical conversations) and assign this norm conversation an average interactivity $\bar{\tau}$ (e.g. "room temperature" $\bar{\tau} = 20°$).

The second complementing assumption is more important and describes the local behaviour of our metric. Suppose that all sojourn times are fixed except for one state *I*. Then, reducing the sojourn time in *I* by a factor of $(1 - \delta)$ must increase the overall interactivity τ by $(1 + \varepsilon)$, say. Hence we get

$$t_I((1 - \delta)\tau; t_{-I}) = (1 + \varepsilon) \cdot t_I(\tau; t_{-I}) \tag{6}$$

for $I \in \sigma$, $\delta \ll 1$ and $\varepsilon = \varepsilon(\delta, \tau) \ll 1$.

Now, standard Taylor expansion of (6) around τ yields the differential equation

$$-\frac{\partial}{\partial \tau} t_I(\tau; t_{-I}) \cdot \delta\tau = \varepsilon(\delta, \tau) \cdot t_I(\tau; t_{-I}). \tag{7}$$

For $\varepsilon(\delta, \tau) = \alpha$ being constant, (7) gives

$$\frac{\partial}{\partial \tau} t_I(\tau; t_{-I}) = -\frac{\alpha}{\delta\tau} \cdot t_I(\tau; t_{-I}), \tag{8}$$

and we can easily verify that the solution is proportional to

$$t_I(\tau; t_{-I}) \propto \frac{1}{\tau}. \tag{9}$$

For the more sophisticated choice $\varepsilon(\delta, \tau) = \delta \cdot \frac{\bar{\tau}}{\tau}$, (7) leads to

$$\frac{\partial}{\partial \tau} t_I(\tau; t_{-I}) = -\frac{\bar{\tau}}{\tau^2} \cdot t_I(\tau; t_{-I}), \tag{10}$$

with solution

$$t_I(\tau; t_{-I}) \propto \exp\left(\frac{\bar{\tau}}{\tau}\right). \tag{11}$$

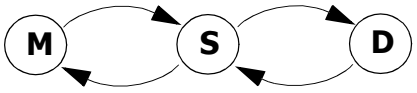

Figure 2: Conversation Model 2 (3 States)

As already discussed in [3], we can interpret Model 1 in two different ways: from a stochastic perspective, it corres-ponds to a continuous-time Markov chain, where the sojourn time in state *I* is exponentially distributed with parameter [10]

$$\lambda_I = \frac{1}{t_I} \propto \exp\left(-\frac{\bar{\tau}}{\tau}\right). \tag{12}$$

Statistical thermodynamics [11], however, allows to interpret the model in terms of a particle jumping over a series of potential walls ΔE, where the particle's success rate at wall *I* equals

$$\lambda_I \propto \exp\left(-\frac{\Delta E}{kT}\right) \tag{13}$$

with Boltzmann's constant *k* and system temperature T. The striking similarity between (12) and (13) leads thus to the interpretation of the interactivity metric τ as "temperature" of the conversation [9].

3 Interactivity-Specific State Models

We start the generalization of the original approach as described in the previous section by taking a closer look at the state model of a conversation. Whereas the 4-state model of P.59 (Model 1) [5] provides a useful way of describing any conversation between two parties for whatever purpose, it is by no means the only option. In terms of interactivity, in a first step it may e.g. be irrelevant to distinguish between actual speakers. Assuming that states "*A*" and "*B*" are symmetric and thus can be merged into a joint state "*S*" (single-talk, i.e. 1 person speaking), we end up with the 3-state model depicted in Figure 2 with additional states "*M*" and "*D*" as before.

On the contrary, for the case of $n \geq 2$ participants it has turned out useful to omit the transition states "*M*" and "*D*" completely. This leads to the *n*-state Model 3 (see Figure 3 for $n = 3$) where all phases of mutual silence and/or double talk are assigned suitably to one of the speakers; therefore Model 3 can easily be extended to an arbitrary number of participants.

Coming back to the case of two participants, we might also be interested in a coarse classification of states according to their "semantic functionality": mutual silence, e.g., may have a significantly different impact on τ, depending on whether state "*M*" describes a "conversation turn" from "*A*" to "*B*" (i.e. A is speaking, and after a pause B is replying), or a "user think time" (i.e. A is speaking, pausing, and speaking again while B remains silent). This is reflected in Figure 4 by introducing two

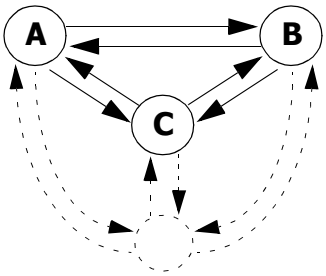

Figure 3: Conversation Model 3 ($n \geq 2$ States)

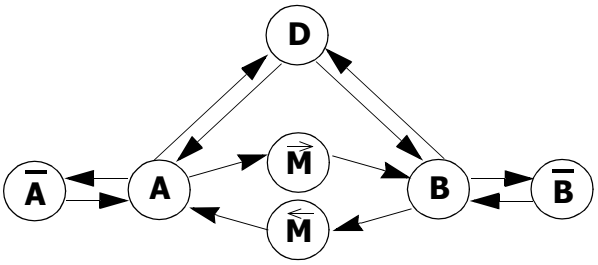

Figure 4: Conversation Model 4 (7 States)

new states "$\bar{A}$" (A is pausing and continuing afterwards) and "$\bar{B}$" (B is pausing and continuing afterwards). Additionally, we have also to split state "M" into "$\vec{M}$" and "$\overleftarrow{M}$", depending on the direction of the transition between "*A*" and "*B*"..

Note that we are only able to fully use the descriptive quality of Model 4 if we are in a position to quantitatively distinguish the differing impact of the individual states on the interactivity τ. To this end, the following section introduces a novel characterization of the conversation states.

4 State-Specific "Heat Capacity"

We have already mentioned that for characterizing conversational interactivity, it may be very interesting to take also the "semantic functionality" of different states into account. There is for instance a clear behavioral distinction between states "*M*" and "*D*" in Model 1: whereas a long period of mutual silence indicates complete absence of interactivity, an equally long period of double-talk indicates an interactivity reduction (due to the limited exchange of information), but not to the same extent as with mutual silence, as there remains still a certain level of information flow as well as reactions to it.

We describe this phenomenon by introducing a new parameter c_I which characterizes each state $I \in \sigma$ in terms of its influence on conversational interactivity: the larger c_I, the larger the impact of extended/abbreviated sojourn times in state *I*. This impact could be measured by user experiments, for instance by observing the "speed" with which the left-hand side of (3) approaches 0 if $t_I \to \infty$ (see Figure 5 for a qualitative illustration).

From our initial discussion of the impact of "*M*" versus "*D*", we may in general conclude that $c_M > c_D$. As for the case of extensive monologues from either speaker A or B, the impact of a long "*A*" or "*B*" period is certainly smaller than the one caused by an equally long period of "*M*", but larger than for "*D*". Summarizing these deliberations for Model 1, this leads immediately to

$$c_D < c_A \approx c_B < c_M \tag{14}$$

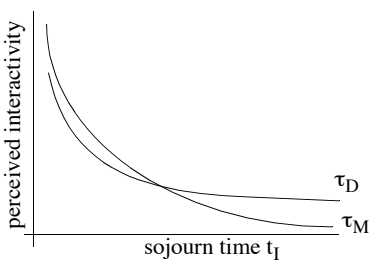

Figure 5: Sojourn Time vs Perceived Interactivity

Determining the exact value of c_I is a non-trivial task and requires careful subjective testing where users assess the perceived interactivity of conversations that vary only in terms of the sojourn time spent in *one* of the states (whereas all other parameters remain fixed). The test results are subject to a non-linear regression with respect to the candidate functions

$$g_h(t_I;t_{-I}) = \frac{1}{c_I \cdot t_I} \quad \text{or} \quad g_e(t_I;t_{-I}) = \frac{1}{c_I \cdot \ln t_I}\,, \tag{15}$$

depending on whether we have chosen the hyperbolic (9) or the exponential (11) temperature model, resp.

As a striking feature of this new parameter, note that there exists again a nice analogy in thermodynamics: there, c_I is well-known as "specific heat capacity" of a given material and describes the amount of energy required for heating 1 kg of the material by 1° C. In the next section, we will use this fact for describing an efficient method to actually calculate the temperature of a conversation.

5 Temperature Calculation

As soon as it comes to the practical calculation of a temperature value for a given conversation, there are two basic interpretations for the solutions of the differential equation (7). [3] and [9] assume the existence of a unique temperature for the complete system while varying average sojourn times in the individual states are a mere expression of statistical fluctuations. Therefore, calculating the conversational temperature here means to determine the unique value of τ for which these fluctuations are minimal (using e.g. a least-square method).

In the present paper, however, we argue for a more straigthforward approach and use the solutions of (7) for assigning an individual temperature $\tau_I = t_I^{-1}(\tau_I;t_{-I})$ to each of the conversation states *I*. The eventual system temperature τ is then calculated as weighted average

$$\tau = \frac{\sum_{I \in \sigma} c_I \cdot \pi_I \cdot \tau_I}{\sum_{I \in \sigma} c_I \cdot \pi_I} \tag{16}$$

as already stated in (1) with $N = \sum c_I \cdot \pi_I$ as normalizing constant, specific heat c_I and mean sojourn probability π_I according to

$$\pi_I = \frac{t_I}{\sum t_I} \quad \text{for states } I \in \sigma\,. \tag{17}$$

Thus, the conversational temperature is calculated as a weighted average of the individual state temperatures, where the weights reflect both the proportion of time during which the conversation remains in the individual states as well as the heuristic impact of the states onto perceived interactivity as expressed by c_I. Note that this is comparable to the analogous thermodynamic problem of determining the equilibrium temperature of a heterogeneous mixture of different amounts (masses) of different materials which initially have different individual temperatures.

The advantage of this approach is obvious: whereas in [3] and [9] a complicated implicit non-linear optimization problem has to be solved, the present approach reduces complexity to a simple inversion of the sojourn-time/temperature relationship in either the hyperbolic form (9) or the exponential form (11) for each of the states *I*.

6 The Multiparty Case

In Figure 3, we have already introduced Model 3 for $n \geq 2$ speakers participating in a joint conversation where the sojourn time span in each of the n associated states $I \in \{A, B, C, \ldots\}$ is defined as the time between the begin of one speaker's talk period and the instant the next one starts talking. Thus, pauses and periods of mutual silence are absorbed into the previous speaker's talk period, and double-talk is considered as a change to the state of the new, intervening speaker.

Having defined the average sojourn times t_I for state $I \in \{A, B, C, \ldots\}$, we can derive the sojourn probabilities π_I according to (17), and the mean sojourn time $\bar{t}$ as

$$\bar{t} = \frac{1}{n} \sum_I t_I \ . \tag{18}$$

In order to extend the conversational temperature metric for this model, we require the following properties:

- The temperature τ increases with decreasing mean sojourn time $\bar{t}$ (see (9) and (11)).
- τ increases with increasing n, i.e. when more speakers participate in a conversation. However, with more and more speakers joining the conversation, this increase should flatten out.
- For given n, τ is maximized if all speakers get an equal share of the conversation (i.e., when the probability distribution π_I is uniform over I).
- If one speaker does not participate at all in a conversation ($\pi_J = 0$ for some J), the temperature should continuously approach the value we get for the same sojourn times, but assuming only $n - 1$ speakers from the beginning.

A solution to the above requirements is provided by

$$\tau \propto \frac{1}{\bar{t}} \cdot \left(-\sum_I \pi_I \log_2 \pi_I \right) \tag{19}$$

where the denominator is equal to the entropy of the distribution π_I. Thus, the entire expression can be interpreted as an entropy rate in the following sense: if we randomly select an observation instant, it takes $-\sum \pi_I \log \pi_I$ bits to decide which speaker is currently active, and this amount of information is required to update our speaker state model every $\bar{t}$ units of time. For a uniform sojourn time probability distribution, the temperature is maximized with $\log_2(n)/\bar{t}$, whereas if on speaker is completely silent all the time, the maximal temperature drops to $\log_2(n-1)/\bar{t}$ as required.

Remember finally that this approach requires a much less sophisticated conversation model with only one state per participant. Figure 6 demonstrates the results of a comparison between conversational temperature (11) (left) and entropy rate (19) (right) as obtained from a series of initial tests with two conversation partners accomplishing five tasks with presumably different interactivities, i.e.: (i) Kitawaki's "Random Number Verification" task [7], (ii) the "weather data exchange" and (iii) the "pizza service" tasks according to [4], (iv) a "free conversation" on given topics, and (v) taking turns in reading a poem. The observed close correlation between the two approaches proves that both metrics are certainly consistent for the case of $n = 2$ participants.

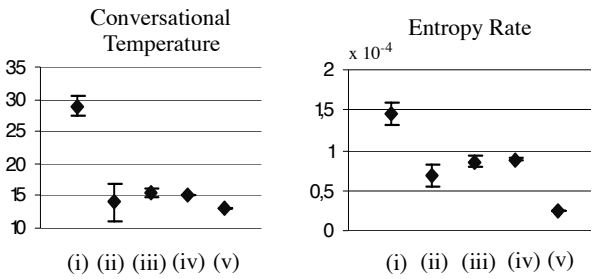

Figure 6: Conversational Temperature vs Entropy Rate

7 Conclusions and Further Work

This paper has sketched the central idea of measuring conversational interactivity with the help of a temperature metric. The original approach described in [9] has been extended towards dealing with several interactivity-specific conversation models, describing the "semantic functionality" of different conversation states with the help of a new parameter ("specific heat"), and investigating conversations with an arbitrary number of participants. The main focus of our current and future work is to apply the concepts developed in this paper to subjective conversation tests. Another focus of future research is directed towards the integration of transmission delay over error-prone links into the framework presented in this paper.

Acknowledgements

Part of this work has been performed in the framework of the Austrian Kplus Competence Center programme.

References

[1] E. Couper-Kuhlen, M. Selling (eds.): *Prosody in conversation*. Cambridge University Press, pp. 11–56. Cambridge (UK), 1996.

[2] F. Hammer, P. Reichl: *How to Measure Interactivity in Telecommunications*. Accepted for: 44th FITCE Congress, Vienna, Austria, Sept. 2005.

[3] F. Hammer, P. Reichl, A. Raake: *The Well-Tempered Conversation. Interactivity, Delay and Perceptual VoIP Quality*. Proc. IEEE ICC 2005, Seoul, Korea, May 2005.

[4] F. Hammer, P. Reichl, A. Raake: *Elements of Interactivity in Telephone Conversations*. Proc. ICSLP 2004, Jeju Island, Korea, Oct 2004.

[5] ITU-T: *Artificial Conversational Speech*. Rec. P.59, March 1993.

[6] ITU-T: *The E-model, a computational model for use in transmission planning*. Rec. G.107, March 2003.

[7] N. Kitawaki, K. Itoh: *Pure delay effects on speech quality in telecommunications*. IEEE JSAC, vol. 9, no. 4, pp. 586– 593, May 1991.

[8] S. Rafaeli: *Interactivity: From new media to communication*. In: *Sage Annual Review of Communication Research: Advancing Communication Science*, vol. 16, pp. 110–134. Beverly Hills, CA, 1988.

[9] P. Reichl, F. Hammer: *Hot Discussion or Frosty Dialogue? Towards a Temperature Metric for Conversational Interactivity*. Proc. ICSLP 2004, Jeju Island, Korea, Oct 2004.

[10] S. M. Ross: *Stochastic Processes*. Wiley 1996.

[11] K. Stowe: *Introduction to Statistical Mechanics and Thermodynamics*. Wiley 1983.