CONVERSATIONAL INTERACTIVITY IN SLOW-MOTION: A COMPARISON OF PHONE DIALOGUES AND OPERA SCENES

Peter Reichl, Florian Hammer
Telecommunications Research Center Vienna (ftw.)
Donaucitystr. 1, A-1220 Vienna, Austria
{reichl; hammer}@ftw.at

Marena Balinova
doremi art management
Seligerstrasse 3, D-91781 Weissenburg, Germany
balinova@gmx.net

ABSTRACT

Modeling the interactivity of conversations recently has gained increasing interest in the telecommunications community, especially with regard to the integration of multimedia applications over packet-based transmission technologies. A quantitative description of interactivity either relies on (subjective) user tests or on instrumental (objective) metrics which use appropriate signal parameters for deriving a scalar characterization of interactivity. Whereas traditional research in this area is based on examples of spontaneous conversations, our main focus is on "non-spontaneous conversations", where structure and speakers' actions are largely fixed a priori, e.g. by a movie script or a music score. As special examples, in this paper we investigate the characteristics of opera duets and larger ensemble scenes with respect to interactivity models like Parametric Conversation Analysis, Conversational Temperature or Conversational Entropy. After introducing the basic measurement framework and reviewing related experiments in the area of VoIP (Voice-over-IP), we present quantitative results for a series of representative opera excerpts. Having demonstrated the close relationship between these two types of conversations, we argue that opera scenes can formally be viewed as conversations in slow-motion. This may lead to important conclusions for telecommunication networks as well as for the design of interactive sound systems and the parametrization of algorithmic composition tools.

1. INTRODUCTION AND MOTIVATION

The integration of multimedia applications over packet-based transmission technologies like the Internet is one of the central challenges for the future success of telecommunications as such. In this context, both the research community and standardization bodies like ETSI or ITU-T have recently shown increasing interest in investigating the impact of conversational interactivity on basic network QoS (Quality-of-Service) parameters like transmission delay or jitter. Unfortunately, it turns out that already *defining the concept* of interactivity is a non-trivial task [5], thus explaining the lack of a universally accepted metric for interactivity.

In order to nevertheless gain a quantitative description, we have e.g. to rely on (subjective) user tests, which are expensive and time-consuming, and which can, in our context, be distinguished into

- *active user tests* where the test persons are actively participating in a conversation which they have to judge/evaluate at the same time;
- *passive user tests* where the test persons observe a (recorded and/or video-taped) conversation between other people.

Alternatively, we could aim at instrumental (objective) approaches which allow to automatically derive a numerical characterization from various signal processing parameters related to a conversation. Recently, the following three approaches for such an instrumental metric have been proposed:

- *Parametric Conversation Analysis (PCA)* which describes a given recorded conversation in terms of the rate of alternation between speakers;
- *Conversational Temperature (CT)* which is based on a state model of the conversation and uses the corresponding mean sojourn times for deriving a scalar interactivity parameter (in analogy to the thermodynamic concept of system temperature);
- *Conversational Entropy (CE)* which calculates the information-theoretic entropy of the conversation and is especially suited for conversations with more than two participants.

However, these approaches mainly differ with respect to the conversation scenarios they are useful for. Figure 1 sketches their main application areas with respect to the duration of the conversation sample and the number of participants. We see that PCA is applicable under the most general circumstances, i.e. independent of the

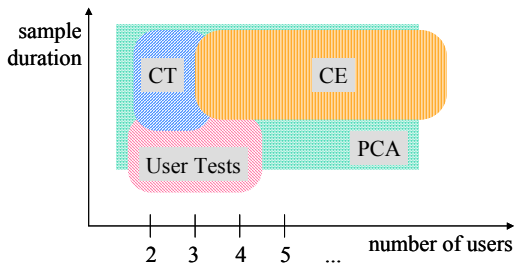

Figure 1: Approaches for Determining Interactivity

number of participants, and delivers best results on a medium time-scale. In contrast, CT is valid only for 2 or 3 participants and long samples, as is CE for a larger number of participants, and user tests are a solid option for short samples with a small number of speakers.
Whereas traditional user tests are based on natural (*spontaneous*) conversation, in [10] we have proposed to supplement this by using also examples of "*non-spontaneous*" conversations which play a vital part e.g. in movies, in theatre, opera, operetta or musical (up to an extent that e.g. Richard Strauss has termed his last masterpiece "Capriccio" no longer an opera but a "conversation piece for music" [13]). We have decided to focus on music rather than movie examples, because in movie dialogues, double-talk episodes are not represented appropriately. As for opera scenes, however, there is a severe lack of analysis concerning their conversational form (cf [1]), and to the best of our knowledge it has never been investigated formally in which way opera duets or larger ensembles are related to ordinary conversations with respect to structure and interactivity. On the other hand, under the fundamental assumption that composers deliberately decide on the structure of their products while trying to mirror the real world as closely as possible, such an analysis might provide us with very interesting insight into conversational details difficult to recover from spoken dialogues. Therefore, this paper proposes some first steps into this direction and provides a couple of interesting initial results.

To this end, we start with extending the basic modeling assumptions for conversational interactivity towards the case of opera duets, leading to a generalized "Conversation/Opera Scenes Model" (COSM) described in Section 2. Section 3. presents the experimental environment and results for speech conversations, before Section 4. deals with quantitative results for selected examples of opera duets, larger ensemblees and video clips. Section 5. closes the paper with some remarks on the lessons learnt.

2. JOINT MODEL FOR THE INTERACTIVITY OF CONVERSATIONS AND OPERA SCENES

This section starts with describing an extended version of the "Parametric Conversation Analysis" due to [4] for the case of opera duets. In a second step, we introduce the concept of "conversational temperature" due to [8], before we briefly touch on a recent entropy-related extension to multi-party conversations like trios and larger ensembles.

2.1. Conversation/Opera Scenes Model (COSM)

Following [6] and [8], we model an arbitrary conversation/ duet between two opera figures using the standard 4-state model depicted in Figure 1. Here, the four states *A*, *B*, *M* ("mutual silence") and *D* ("double-talk") depend on whether speaker/singer 1 and/or 2 is active or silent (see Table 1).

	# 1 silent	# 1 active
# 2 silent	M (E)	A
# 2 active	B	D (E)

Table 1: Conversation/Opera Scenes Model (COSM) States

Note that, in contrast to [8], for our case we cannot omit the direct transitions between *A* and *B* (dashed line), because this switching event is not uncommon in opera scores, whereas it hardly plays a role in everyday conversations. Moreover, we need to introduce an extra state (connected with all other states by dotted lines), i.e. *E* for exceptions (including chorus or secondary characters interrupting briefly as well as extensive periods of parallel singing, i.e. repeated direct transitions between *M* and *D*). These extensions, however, have no futher impact on the validity of the model in general.

2.2. Generalized Framework of Parametric Conversation Analysis

Inspired by the classical study [2], in [4] we have defined a couple of "conversational events" as being relevant for characterizing an arbitrary conversation in detail. The following list presents an adapted version, taking the additional direct transitions $A \leftrightarrow B$ and $M \leftrightarrow D$ into account:

- *Speaker Alternation:* switching event from one speaker to the other, potentially separated by a period of mutual silence. This includes all transitions $A \rightarrow B$, $A \rightarrow M \rightarrow B$, $B \rightarrow A$ and $B \rightarrow M \rightarrow A$ [1].
- *Interruption:* one speaker interrupted by the other who after a phase of double-talk eventually gains

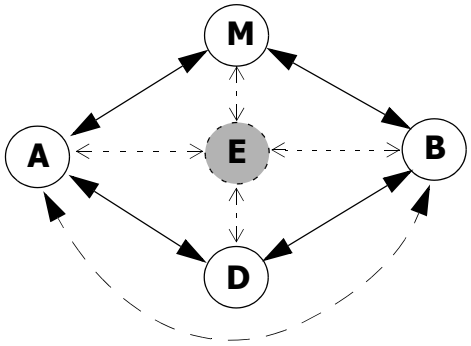

Figure 2: General COSM

the floor, i.e. transitions $A \to D \to B$ and $B \to D \to A$ [2].

- *Pause:* mutual silence phase between two talk-spurts of the same speaker, i.e. $A \to M \to A$ and $B \to M \to B$.
- *Incision (= "non-interruptive double-talk" [3]):* (short) double-talk phase without speaker alternation, i.e. $A \to D \to A$ and $B \to D \to B$.
- *Parallelism:* joint switching of both participants between mutual-silence and double-talk phases, i.e. $M \to D$ and $D \to M$.

Note that except for the last event which is more or less unique to the operatic case, all other events also appear during ordinary conversations (see [4]). On the other hand, there is hardly any opera duet which renounces to include large episodes of parallelism as a simple mean to express simultaneously the thoughts and emotions of the involved persons through a parallel (or even unisono) musical evolution. In fact, traditional duets generally consist of (short or longer) sequences of solo phrases (speaker alternation) switching to extended phrases of parallel singing[3].

In order to make conversations and duets formally comparable, it is therefore necessary to find a way for eliminating this opera-specific feature (which, as already explained earlier, is included within the "exception state" E). We propose therefore to simply omit all episodes E and investigate only the remaining body of the scene/conversation. This of course implies that many scenes, especially most of the examples mentioned so far, could turn out to be useless for our purposes, if they contain e.g. too extended periods of parallelism. Fortunately enough, this is not the case for all opera duets, as we will demonstrate in Section 4.

In any case, according to [4], there are two basic parameters describing its interactivity:

- *Speaker Alternation Rate (SAR):* total number of speaker alternations plus interruptions per minute
- *Interruption Rate (IR):* total number of interruptions per minute.

1. For typical examples we refer to the love duet "*Già nella notte densa*" in Verdi's Otello or the central arguement between Tosca and Scarpia "*Ed or fra noi parliam da buon amici*" in the second act of Puccini's Tosca.
2. To mention just one particular example where this pattern is used extensively, see e.g. the first half of the flower duet "*Scuoti quella fronda*" (Puccini, Madama Butterfly).
3. Note that this typical structure relates to a vast range of famous examples, starting from "*La ci darem la mano ... Andiam, andiam, mio bene*" (Mozart, Don Giovanni) and "*Mira, Norma*" (Bellini, Norma) over "*È lui! desso! l'infante! ... Dio, che nell' alma infondere*" (Verdi, Don Carlo), "*Au fond du temple saint*" (Bizet, Pecheurs des Perles) or "*Oh sink hernieder, Nacht der Liebe*" (Wagner, Tristan) up to duets like the famous barcarolle "*Belle nuit, o nuit d'amour*" (Offenbach, Contes d'Hoffmann), "*Bess, you is my woman now*" (Gershwin, Porgy and Bess) and "*No more talk of darkness*" (Webber, Phantom of the Opera), to name but a few.

2.3. Conversational Temperature Metric

A second approach to formally describe the interactivity of conversations starts from the fact that even if it is hard to describe what interactivity is, there is an intuitive knowledge *what it is not.* In this sense, as pointed out in [8] in great detail, the lack of interactivity is (strongly) correlated with long (infinite) sojourn times in one of the COSM states A, B, M or D. Note that there is a striking analogy in statistical thermodynamics[4], where the activity of particle movements is described in a very simple way by just one scalar parameter, i.e. the system temperature. Interpreting Figure 2 as a continuous-time Markov chain, in [8] and [4] we have therefore derived formally a temperature metric for conversational interactivity. This requires only to measure the mean state sojourn times t_A, t_B, t_M, t_D and compare them to the respective mean values of an appropriate "reference conversation" (typically averaged over a large set of representative conversation recordings and related to by a reference temperature $\bar{\tau}$). The resulting "*conversational temperature*" τ is then calculated as a least-square fit (weighted by the state probabilities) for

$$\lambda_I = \frac{1}{t_I} \propto \exp\left(-\frac{\bar{\tau}}{\tau}\right) \text{ for states } I = A, B, M, D, \quad (1)$$

which closely resembles the famous "Boltzmann's statistic" in classical thermodynamics

$$\lambda_I \propto \exp\left(-\frac{\Delta E}{kT}\right), \quad (2)$$

where ΔE refers to the difference of energy potentials, k is known as Boltzmann's constant and T corresponds to the temperature of the system [12].

As an intuitive illustration of this approach for the case of artificially generated speech, Figure 3 sketches the resulting talk-spurt pattern for one-minute conversations at different temperatures $\tau = 5°, 20°, 40°$,

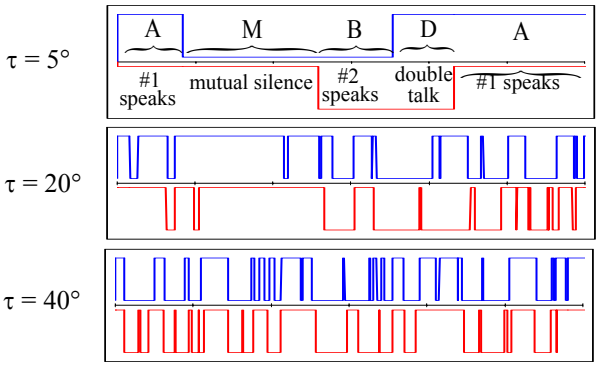

Figure 3: Conversational Temperature: Artificial Speech Examples taken from [8]

4. Note that of course it is not by chance that the presence and/or absence of interactivity often is characterized in terms of a temperature, a "hot discussion" being the most notable example!

where speaker 1 activity is depicted above the x-axis, speaker 2 below, and the x-axis itself corresponds to periods of silence. Note that the reference conversation is assigned "room temperature", i.e. $\bar{\tau} = 20°$, whereas the other examples correspond to temperatures of 5° and 40°, respectively. Later on, in Figure 5, we will present similar diagrams for a couple of selected opera duets.

2.4. Entropy-based Approach

As a last alternative for an interactivity metric, we refer to the entropy-based approach proposed in [9] . The "Conversational Entropy" (CE) is especially appropriate for conversations with more than two participants. Here, we start from a much simpler COSM with only one state per participant, where each sojourn time starts with the respective participant becoming active, and ends with the next participant becoming active, irrespective of the behaviour of the former one. This defines a probability distribution π_I for the various states I of the COSM whose information-theoretic entropy can be calculated according to

$$\frac{-\sum_I \pi_I \log_2 \pi_I}{\frac{1}{n}\sum_I t_I} \qquad (3)$$

and allows another appropriate characterization of the interactivity.

3. EXPERIMENTAL RESULTS FOR SPEECH CONVERSATIONS

3.1. Experimental Setup and Measurement Framework

In a first step, we have validated our approach through a series of conversation experiments performed at the Institute of Communications Acoustics (IKA) at Ruhr-University Bochum, Germany. Here, test persons had to perform different tasks, i.e. either

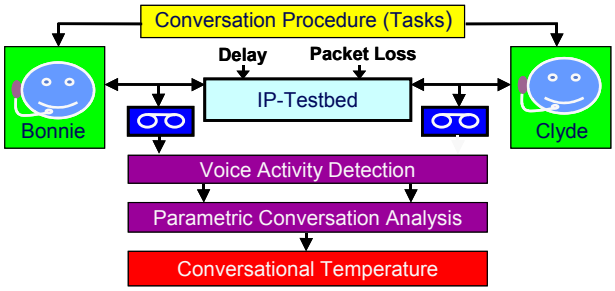

Figure 4: Experimental Setup and Measurement Framework

so-called "short conversational tests" like the ordering of pizzas or the booking of flights, or so-called "interactive short conversational tests", including the fast exchange of meteorological data or telephone numbers. The second type of tasks has deliberately been designed in order to show a significantly different interactivity of the resulting conversations. For each scenario, the conversations of 10 pairs of test persons have been recorded and evaluated. The talk spurts have been determined automatically, under the additional rule that a pause after a talk spurt has to exceed 200 ms in order to be accepted as a state of its own (otherwise, the conversation is assumed to remain in the previous state) [6]. All conversation tests have been conducted in German. Figure 4 sketches the setup of our experiments.

3.2. Measurement Results

The ITU-T recommendation P.59 [6] contains standardized average numerical values for state sojourn times which are derived from a large set of conversations in Italian, English and Japanese. Therefore, in a first step we have used these parameters for our "reference conversation" (with temperature 20°) and have compared them to the results of our conversation tests.

The results of the Parametric Conversation Analysis of Section 2.2. for our user experiments as well as the corresponding reference values for P.59 are summarized in Table 2. The detailed evaluation performed in [3] concludes that the state probability for the mutual-silence state may serve as an important indicator for increasing interactivity, as its value in the pizza scenario is significantly larger than for the more interactive scenario 2 (weather). Overall of course, both SAR and IR clearly indicate the higher interactivity of the weather scenario.

Task / State	A	B	M	D	SAR	IR
Pizza Service	34.3	34.8	27.3	3.6	19.66	4.28
Weather Data	37.8	38.7	18.6	4.9	26.04	5.91
Reference: Rec. P.59	35.2	35.2	22.5	6.6	n.a.	n.a.

Table 2: State Probabilities [%] for Scenarios 1 and 2 and Corresponding SAR and IR

Table 3 presents the average sojourn times in the four states of the COSM (Section 2.1.) for both scenarios (pizza and weather) as well as the reference values from P.59, and finally also the resulting "conversational temperature" according to Section 2.3. (cf. [3]). First of all, we notice that the sojourn times in our experiments are nearly throughout significantly longer

than in the reference case (which might be a particularity of the German language). Therefore, it is not astonishing that the conversational temperature in the test conversations are much lower (around 14°) than in the reference case (20°). However, as expected, scenario 2 (exchange of weather data) turns out to be significantly more interactive than scenario 1 (pizza service), which fact is reflected both in shorter average sojourn times as well as in a higher corresponding temperature.

Task / State	A	B	M	D	Temperature
Pizza Service	1.45	1.59	0.68	0.35	13.4°
Weather Data	1.44	1.44	0.42	0.33	14.4°
Reference: Rec. P.59	0.78	0.78	0.51	0.23	20.0°

Table 3: Sojourn Times for Scenarios 1 and 2 and Resulting Conversational Temperatures

In the next section, we will determine corresponding parameter values for selected opera duets and afterwards use the above results to determine whether and to which extent opera duets are structurally compatible with ordinary telephone conversations.

4. MEASUREMENT RESULTS FOR OPERA SCENES

4.1. Example Selection Criteria

In this section, we evaluate a couple of carefully selected opera scenes using the above models. We choose as the main selection criterion that, at least to the naive listener, the scenes should resemble conventional (everyday) conversations as closely as possible. Moreover, the scenes should

- be representative for different composition forms, different languages and different epoches/styles, preferably taken from well-known works authored by famous composers;
- be representative for different emotional content and communication style, leading to apparently highly interactive scenes;
- have a relatively small probability for state *E* (thus avoiding the parallelism problem, see Section 2.2.).

The appendix contains a detailed description of the selected examples including references to the CD (and/or video) recordings used for our evaluations. Finding good examples fulfilling the listed criteria is not easy, but nevertheless, our collection includes

- duets and ensembles in the form of recitatives, full-fledged duets/ensembles and dramatic scenes;

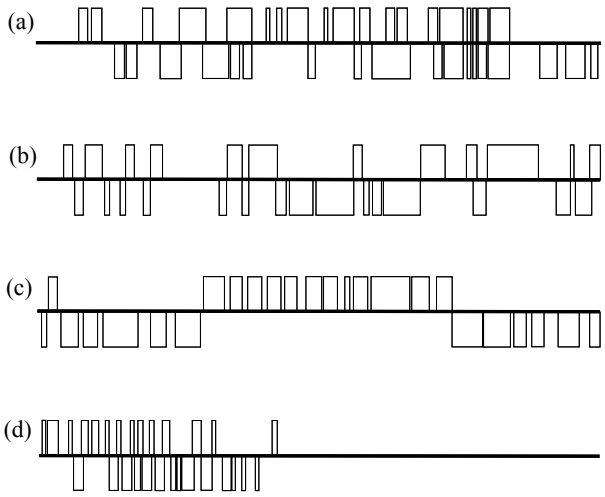

Figure 5: Example Duet Patterns (first 120 secs each): (a) Fidelio (b) Cavalleria Rusticana (c) Carmen (d) Tosca

- examples from the 18th to the 20th century, including Belcanto, Viennese classic, verismo, grand opéra and opéra comique as well as Wagner, Verdi and Strauss;
- examples in Italian, German and French (i.e. the main three opera languages);
- examples ranging from opera buffa to melodramma and tragic opera;
- scenes that deliberately do not contain longer episodes of explicit parallelism/simultaneous singing or external interruptions (choruses etc).

4.2. Scenario 1: Duets

First, we have focussed on duet examples (see Appendix A.1) and have applied the models *without any adaptation* (retaining, e.g., ITU-T's 200 ms rule mentioned in Section 3.1.). Figure 5 presents a couple of graphical illustrations of the initial two minutes of some examples, whereas Table 4 and Table 5 contain numerical values of parameters like state probabilities, alternation rates, mean sojourn times, temperatures etc.

Task / State	A	B	M	D	SAR	IR
Fidelio	22.9	30.9	22.7	23.4	9.66	4.59
Cavalleria	18.9	41.3	21.4	18.5	7.49	2.19
Carmen	41.5	26.7	22.6	9.2	4.20	1.08
Tosca	24.0	37.8	31.0	7.2	21.15	9.40
Duet Averages	31.0	31.0	24.4	14.6	10.63	4.32
Convers. Averages	36.4	36.4	23.0	4.3	22.85	5.10

Table 4: State Probabilities [%] for Duet Examples and Corresponding SAR and IR

Task / State	A	B	M	D	CT
Fidelio	1.53 (0.99)	2.06 (1.21)	1.21 (1.00)	1.56 (0.90)	9.6° (12.8°)
Cavalleria	2.43 (1.82)	3.72 (2.12)	1.67 (0.66)	2.64 (1.22)	7.4° (9.7°)
Carmen	3.28 (2.86)	3.04 (2.70)	1.24 (0.68)	3.51 (3.46)	7.3° (7.9°)
Tosca	0.87 (0.73)	1.36 (1.45)	0.92 (0.39)	0.45 (0.47)	13.3° (14.8°)
Average Convers.	1.48 (n.a.)	1.48 (n.a.)	0.50 (n.a.)	0.34 (n.a.)	12.0°

Table 5: Mean Sojourn Times for Duet Examples (medians in brackets) and Resulting Conversational Temperatures

4.3. Comments and Model Adaptation

The last two lines of Table 4 includes the average state probabilities of the selected couple of duets and the respective figures from our own speech conversation tests (Table 2). Already based on this small selection of four duets, we can assume that in all three cases, i.e. ITU-T Rec. P.59, our own speech conversation tests and the new results from opera duets, there is roughly a 1 : 1 : 1 relationship between the total time spent in state *A*, state *B* and combined states *M+D*. Hence, for about one third of the whole duration, speaker A is active, the second third goes to speaker B, and the last third is spent in mutual silence or double-talk. This „*rule of thirds*“ is even more evident in Table 6 where we compare a larger set of samples. Thus, as a first fundamental insight, we note that with respect to state distributions, opera duets and speech conversations are more or less undistinguishable.

Task / State	A	B	M	D	SAR	IR
Barbiere	54.6	30.8	13.6	1.0	12.3	0.8
Fidelio	26.0	33.5	22.8	17.8	8.7	4.0
Cavalleria	31.3	41.8	19.1	7.8	8.3	3.0
Carmen	43.1	22.4	34.5	0.0	4.5	0.0
Walküre	14.1	43.9	42.0	0.0	2.8	0.0
Otello	42.7	35.4	19.9	2.0	10.5	2.9
Tosca	25.0	39.4	28.4	7.2	21.4	9.5
Average	34.6	34.6	26.4	7.2	9.8	2.9

Table 6: State Probabilities [%] for Duet Examples and Corresponding SAR and IR (Revised Scenario 1)

Moreover, already from the very limited set of experiments included in Tables 4 and 5 we can draw a couple of interesting conclusions. The first of those is related to the mentioned 200 ms rule of P.59 (a mutual silent period is detected only if it lasts longer than 200 ms). Whereas this rule has been proposed for the context of speech analysis, according to our experience it is no longer useful for our purposes, as it often leads to classifying a semantic unit which is interrupted by a small pause as two different units with a mutual silence period in between. Whereas it is not easy to derive a more appropriate period based on speech conversations, music makes life much easier, as it is most helpful to consider in detail the musical phrasing of our examples. Here, usually any pause below 500 ms is spanned by some melodic support, and only above this limit, a new musical idea may appear. Therefore, 500 ms seems to be a much more appropriate limit for the a minimal pause duration and is used in the rest of the paper.

Secondly, as far as the mean sojourn times are concerned (Table 5), in some cases we observe a significant difference between the average values and the medians. This difference is caused by the existence of singular periods of very long sojourn times in one state (e.g. orchestra interludes for state *M* or long phrases in states *A*, *B* or *D*). On the other hand, the median somehow avoids this bias and thus provides an important alternative to be investigated later on. Thus, for the rest of the paper we will provide results for both the average and the median option.

4.4. Scenario 1 Revisited

Based on the remarks made in Section 4.3., we have used an evaluation method based on a limit of 500 ms and a larger set of samples. The results for the average state probabilities are depicted in the last line of Table 6 and now reveal a really exiting coincidence between the speech and opera samples in the sense of the „rule of thirds“.

Task / State	A	B	M	D	CT
Barbiere	2.94 (1.59)	2.33 (1.61)	0.64 (0.57)	0.78 (0.78)	26.7° (29.0°)
Fidelio	1.93 (1.09)	3.48 (2.30)	1.58 (1.31)	1.85 (0.69)	23.7° (23.3°)
Cavalleria	3.78 (2.21)	5.05 (2.57)	1.98 (0.69)	1.41 (0.83)	15.2° (17.7°)
Carmen	4.27 (2.93)	3.16 (2.63)	2.20 (0.79)	n.a.	17.1° (15.5°)
Walküre	2.81 (1.83)	6.71 (4.89)	3.76 (1.41)	n.a.	12.1° (11.5°)
Otello	2.52 (1.83)	3.90 (2.32)	1.18 (0.67)	0.36 (0.27)	22.0° (21.0°)
Tosca	0.97 (0.75)	1.81 (1.75)	1.10 (1.07)	0.45 (0.47)	107.6° (34.0°)
Average	3.49 (2.22)	3.49 (2.22)	1.90 (0.88)	1.16 (0.68)	20.0°

Table 7: Mean Sojourn Times for Duet Examples (medians in brackets) and Resulting Conversational Temperatures CT (Revised Scenario 1)

Table 7 provides the mean sojourn times for the revised duet scenario. Note that the average values here are approximately 2-3 times larger than for our speech experiments and 5 times larger than P.59. As far as the median values are concerned, they are only 1.5 times larger than the speech experiments and about 3 times larger than P.59 (except for state *M*).

Therefore, summarizing Tables 6 and 7 into the fundamental proposition of this paper, we can interpret opera duets as a form of "*natural speech conversations in slow-motion*", i.e. with the same overall relationship between the speakers, but on a larger time-scale. As there is no doubt on the usefulness of visual slow-motion for certain types of research, our approach of „*acoustical slow-motion*" might prove equally interesting e.g. for the investigation of speech conversations, but on the other hand will have further applications also in the area of interactive sound systems and the parametrization of algorithmic composition tools.

4.5. Scenario 2: Larger Ensembles

In a next step, we investigate larger ensembles, i.e. with more than 2 participants. Appendix A.2 lists our selection containing three trios, one quartet and two large scenes (with up to 8 singers each, who are now represented by states *S*, *T*, *U*, *V*, *W*, *X*, *Y* and *Z*)[5].

Task / State	S [W]	T [X]	U [Y]	V [Z]	SAR
Cosi fan tutte	24.8	22.3	52.9	n.a.	19.4
Magic Flute	40.2	35.7	24.1	n.a.	16.9
Faust	35.0	46.2	18.8	n.a.	11.2
Rigoletto	54.5	13.7	19.6	12.2	12.6
Salome	10.5 [10.3]	11.3 [12.3]	32.2 [11.4]	9.3 [2.8]	20.8
Gianni Schicchi	18.3§ [15.3]	4.2 [13.4]	5.3 [3.9]	25.5 [14.0]	15.0

Table 8: State Probabilities [%] for Ensembles and Corresponding SAR

Tables 8 and 9 correspond exactly to Tables 6 and 7 and contain state probabilities and mean sojourn times as well as SAR and resulting entropy. Note that we omit IR and focus on Conversational Entropy (CE) rather than on Conversational Temperature even if the latter could in principle be calculated at least for the case of $n = 3$ participants as well (with a 9-state-model replacing Figure 2). The calculation of averages does not make sense any more and is omitted.

5. In the last two examples (Salome and Gianni Schicchi), the eight states have been distributed pairwise into the 4 table cells in order to maintain readability. Thus, the last four states appear in square brackets.

Task / State	S [W]	T [X]	U [Y]	V [Z]	CE
Cosi fan tutte	2.07 (1.67)	2.20 (2.07)	5.73 (3.93)	n.a.	0.42
Magic Flute	4.28 (3.55)	3.57 (2.24)	2.93 (2.27)	n.a.	0.44
Faust	6.98 (4.45)	7.07 (2.53)	2.88 (1.87)	n.a.	0.26
Rigoletto	7.70 (2.49)	2.83 (1.56)	3.05 (2.69)	5.43 (3.72)	0.31
Salome	3.58 (3.19) [2.64] (1.22)	5.81 (5.59) [1.81] (1.19)	4.71 (2.63) [1.79] (0.84)	1.90 (1.32) [5.72] (5.72)	0.81
Gianni Schicchi	3.92 (3.55) [6.56] (5.51)	2.70 (2.70) [4.31] (4.23)	1.72 (1.89) [5.03] (5.03)	5.46 (2.95) [3.59] (2.87)	0.70

Table 9: Mean Sojourn Times for Ensembles (medians in brackets) and Resulting Conversational Entropy (CE)

4.6. Preliminary User Tests (Audio and Video)

Finally, we have also conducted initial user tests to further validate our results. To this end, we have asked six users (male and female, musicians as well as non-musicians, German and Italian) to judge 8 duet samples, 8 ensembles and 8 video clips with durations between 25 and 60 secs each. Users have been calibrated by exposing them to two episodes from Fidelio (average interactivity) as well as one of Walküre (very low activity). The test samples have then been taken from the rest of the collection described in the Appendices, each opera being represented once or twice. The basic idea behind the choice of short samples was to preserve homogenity, as it turned out that longer episodes often provide a certain variation of the perceived interactivity and thus increase the uncertainty of the test users.After the calibration procedure, users were asked to listen to the 8 duet samples (in random order, i.e. changing for every test user) and to indicate their perceived interactivity by making a cross along a line of length 5 (where 0 represents very low interactivity and 5 very high interactivity, see Figure 6).

In each case, the users were also asked to indicate whether determining the perceived interactivity of the sample was a hard task for them or not. This was followed by eight ensemble examples which very evalu-

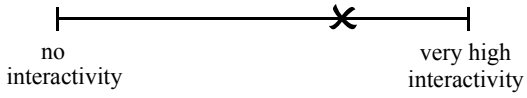

Figure 6: Interactivity Scale for User Tests

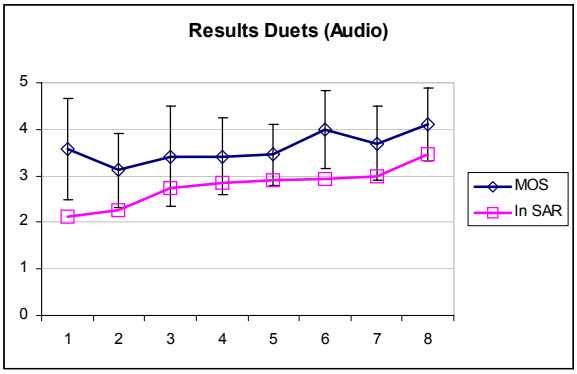

Figure 7: User Test Results I: Duets, Listening-only

ated exactly the same way. Finally, the eight initial duet examples were repeated, but now in the form of excerpts from opera movies (i.e. no video-taped stage performances, but dedicated movies by famous directors like Zefirelli or Ponnelle). Here again, the tasks were identical to the listening-only test. In total, each user test took roughly 45 mins.

Figure 7 presents the duet results (MOS = Mean Opinion Score) versus logarithm of SAR, in ascending order according to the SAR of the samples, showing also the standard deviations of the evaluation results. Except for sample 1 where we observe a significant difference between users' estimation and SAR, the rest demonstrates a reasonable relationship between the two parameters, the cross-correlation between MOS and SAR being 0.76 (and even 0.86 if we exclude sample 1 from the calculation).

In Figure 8, we provide the test results for the eight samples representing larger scenes (ensembles). Here, the correlation between SAR and MOS reaches 0.93 which is quite high. The only outlying example is the introduction scene from Gianni Schicchi which consists of short utterances separated by long pauses, thus indicating that user tests may become less helpful for the case of many participants (see Figure 1).

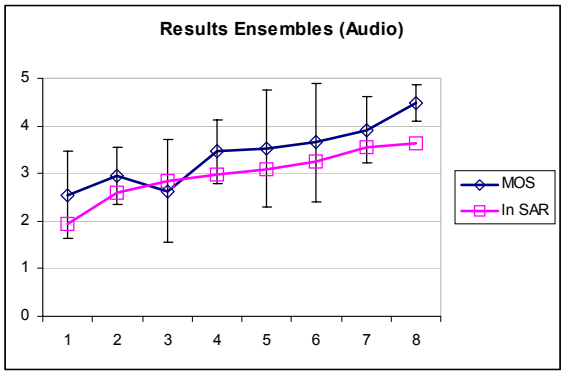

Figure 8: User Test Results II: Ensembles, Listening-only

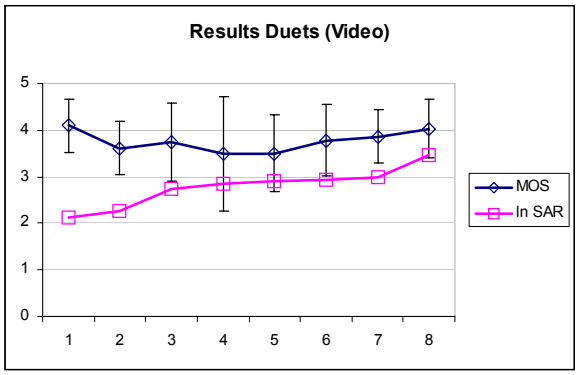

Figure 9: User Test Results III: Duets, Video

Finally, Figure 9 shows the user evaluation for the video clip collection of the eight duets investigated at the start of the experiment. In contrast to Figure 7, there is no clear trend to be identified, mainly due to the first three examples (especially example 1 is outlying again). The overall correlation is going down to 0.17 (but is still 0.69 without if example1 is omitted).

4.7. Discussion

For our experimental evaluation, we have chosen a wide range of examples, structures and patterns, indicated by our selection criteria as well as by the pattern sketch of Figure 5. Thus, despite the obvious divergences within the example set, it is more than astonishing that in the case of duets, the overall average values for state probabilities follow closely the „rule of thirds" derived from speech conversations (see Table 6 compared to Table 2). Together with the state sojourn times being 3-5 times longer for duets than for speech conversations (see Table 7 compared to Table 5), this has led us to the fundamental conclusion that opera duets are a form of conventional speech dialogues in slow-motion. Interestingly enough, we could also interpret the pause duration limit of 500 ms we found in opera scores as being roughly 3 times as large as the value proposed in P.59 for artificial speech generation, however we firmly believe that also for speech conversation, a limit of 500 ms is more appropriate to distinguish semantic units. This is also subject to further research.

The graphical interpretation of Figure 5 allows already a provisional ranking of the examples with respect to their interactivity as later confirmed by user tests. Applying the interactivitiy models to the opera examples without any adaptation results in characterizing them as having quite low interactivity: Table 4 shows that the SAR for duets is only half the value for speech conversation, the IR being approximately equal. This is also supported by the CT in Table 5 and

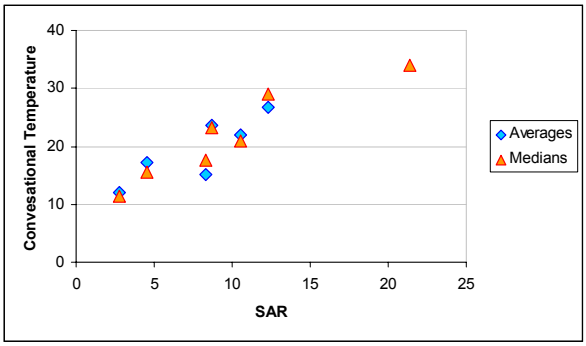

Figure 10: CT versus SAR (Duets)

of course is fully in line with our „slow-motion“ characterization of opera scenes.

Another problem with our initial scenario is related to fluctuating interactivity already within a scene of more than one minute duration. Therefore, we have redesigned our examples in order to maintain homogeneity within each clip and have also applied a pause limit of 500 ms. Moreover, we have chosen the average state sojourn times of our samples as characteristic for the „norm duet“ which has been assigned „room temperature“ of 20° in the CT approach. Under these adaptations, both instrumental interactivity metrics (SAR, CT) are highly correlated and provide an identical ranking among the examples (see Tables 6 and 7).

For our evaluation of scenes with more than 2 participants, we have focused on the entropy approach (Table 9). Compared to the SAR (Table 8), we again note a high correlation and identical ranking of the samples. This proves that SAR may provide a sufficient link between all three interactivity metrics.

Moreover, the listening-only tests presented in Section 4.6. indicate that there is also a strong link between SAR and user perception. As a quite interesting fact, we have seen that this seems to be even more obvious for the logarithm of the SAR, which somehow reminds us of the Weber-Fechner law which e.g. has led to the dB metric of acoustical loudness. However,

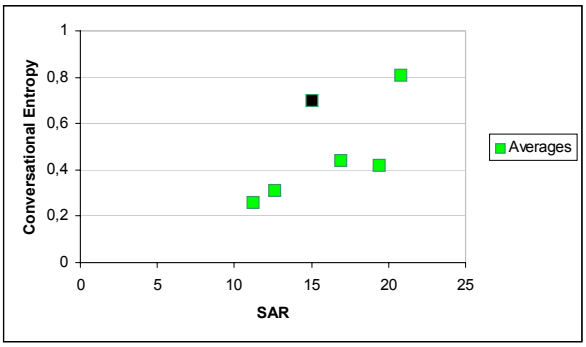

Figure 11: CE versus SAR (Ensembles)

as far as listening-only tests are concerned, some of the examples have turned out to be definite outliers, especially for small values of SAR. This is even more striking in the case of video clips of the same scenes, where the examples with high SAR are evaluated rather independently of the form of the test, whereas for smaller SAR, the video test results are not satisfying. This may have a variety of reasons, like the cut frequency or the emotional content of the scenes as expressed by physical actions (e.g. the fight between Carmen and Don José has received a much higher user evaluation in the video case than for listening-only).

The other apparent outlier example is the introduction of Gianni Schicchi. Here, there is both a discrepancy between SAR and the user evaluation (Figure 8 example 3) and between SAR and CE (Figure 11, third point from left). One possible explanation is related to the fact that this example includes an unusual high percentage of mutual silence. Note of course that these are only preliminary remarks, and there is much room left for further research and interpretation of such phenomena.

All in all, Figures 10 and 11 provide a final summary of our results. From Figure 10 we learn that except for the CT of Tosca, which reaches 107° in the case of average sojourn times, for both average and median cases of Table 7, CT and SAR are highly correlated (correlation coefficient $\rho = 0.90$ for the averages and even $\rho = 0.94$ for the case of medians). Similarly, Figure 11 demonstrates the correlation $\rho = 0.67$ between CE and SAR for the ensemble case, which rises up to $\rho = 0.83$ if we exclude Gianni Schicchi (black dot) from the calculation.

4.8. Further Application of the Results: Algorithmic Composition

Our concept of measuring the conversational interactivity cannot only be applied to analyse opera pieces and telephone conversations. Since the conversational temperature metric is based on a continuous-time Markov chain, it may be used as a tool for algorithmic composition [11]. Different temperatures and timescales result in sound synthesis control patterns not only for event on-/off-sets, but also pitch and spatial information. This promising application is of course out of the scope of this paper, but of course subject of potential future work.

5. SUMMARY AND CONCLUSIONS

This paper has introduced several metrics for describing conversational interactivity and has presented first results demonstrating the expressiveness of the approaches for the case of phone conversations.

The models have been generalized to cope with specific features of non-spontaneous conversations, opera scenes being the most notable example for this class of conversations. Using several representative examples, we have proved the applicability of our methods also for this case. It is especially remarkable that all instrumental metrics are highly correlated among themselves as well as with respect to the results of the user tests we have performed, at least in the case of listening-only tests.

Thus, we conclude that our approaches allow for a purely technical description of conversational interactivity providing interesting insight in various different communication scenarios. Additionally, we have learnt several lessons from this investigation, among them the following: (1) the role of the 200 ms requirement for identifying an M-state is obviously based on a different technical rationale and should most probabtly be revised for the context of interactivity; (2) the different states need appropriate weighting in order to handle e.g. the parallelism problem; (3) the role of emotional content for user perception needs to be clarified. Nevertheless we conclude that, as soon as future research has solved these issues, non-spontaneous examples will be able to provide a significant contribution to further investigate conversational interactivity.

Note finally that the results of the above analysis provide valuable information for the choice of parameters in music composition, too. Thus, our interactivity metrics might also be a useful parameter for the design of interactive sound systems and algorithmic composition.

ACKNOWLEDGEMENTS

This work has been performed partially in the framework of the Austrian Kplus Competence Center programme. The authors want to thank Gernot Kubin and Alexander Raake for helpful discussions and further support.

APPENDIX: DETAILED DESCRIPTION OF TEST SAMPLES

This appendix contains a concise description of the recordings used in the evaluation section of the paper, at the same time indicating what the test users have been told about the rough content of the scenes before being exposed to them. Note that the duration times refer to the the part presented in the user tests, whereas the initial comparison with the ITU-T rec. P.59 has been performed with respect to the whole length of the duets in order to take advantage of the additional material.

A.1 Duets

- **Gioacchino Rossini, Il Barbiere di Siviglia, Act I: Recitativo Figaro - Rosina**

We start our experimental investigation by a detailed look to what we expect to be the musical form closest to everyday conversation, i.e. the "recitativo". Our example is taken from the first act of Rossini's "Barber of Seville". Figaro (A) meets Rosina (B) and reports that Count Almaviva is in love with her ("*Ma bravi! ma benone!*", duration approx. 2:35 mins. CD recording used: Gruberova-Chernov/Weikert 1997; video: Berganza-Prey/Abbado).

- **Ludwig van Beethoven, Fidelio, Act I: Duet Marzelline - Jaquino**

Our next example is taken from the introduction to Beethoven's "Fidelio". Shy Jacquino (A) wants to propose to Marzelline (B) but is repeatedly disturbed by strangers knocking at the door ("*Jetzt, Schätzchen, jetzt sind wir allein*", duration approx. 3:40 mins. CD recordings: Laufkötter-Farell/Walter 1941; Dallapozza-Popp/Bernstein)

- **Pietro Mascagni, Cavalleria Rusticana: Duet Santuzza - Turiddu**

As another example, we investigate the duet "*Tu qui, Santuzza? - Qui t'aspettavo*" from "Cavalleria Rusticana". The scene comprises a highly emotional argument between Sicilian peasant Turiddu (A) and his girlfriend Santuzza (B) who accuses him of having an affair with his former love Lola (duration approx. 3:45 mins until Lola's song. CD recording: Del Monaco-Simionato/Serafin 1951; video: Obraztsova-Domingo/Prêtre).

- **Georges Bizet, Carmen, Act IV: Final Duet Carmen - Don José**

A scene of similar emotional involvement is the end of the final duet of Bizet's "Carmen", where the ex-soldier Don José (A) stabs Carmen (B), his former love who has left him for torero Escamillo ("*C'est toi? - c'est moi!*"), where the evaluation starts from "*Tu ne m'aimes donc plus?*", duration approx. 4:35 mins. CD recording: Baltsa-Carreras/Karajan 1983; video: Bumbry-Vickers/Karajan).

- **Richard Wagner, Die Walküre, 2nd Aufzug: Scene Fricka - Wotan**

Wotan (A), presiding god at Walhall, is blackmailed by his wife Fricka (B) who forces him to approve the assassination of his son Siegmund ("*Was verlangst Du? - Lass von dem Wälsung!*", duration approx. 4:25 mins. CD recording: Meier-Morris/Haitink 1988).

- **Giuseppe Verdi, Otello, Act III: Scene Otello - Desdemona**

In this scene, the jealous moor Otello (A) discusses with his wife Desdemona (B) whom he believes to have an affair with officer Cassio ("*Ma riparlar vi debbo di Cassio*", duration approx. 3:00 mins. CD recording: Scotto-Domingo/Levine 1978; video: Ricciarelli-Domingo/Maazel).

- **Giacomo Puccini, Tosca, Act II: Finale**

Finally, we evaluate a short scene taken from the famous Act II final of Puccini's "Tosca" in order to see the probable upper limit of interactivity to be reached by opera. The brutal police officer Scarpia (A) is killed by famous singer Floria Tosca (B) as he attempts to rape her ("*Tosca, finalmente mia! - Maledetta!!*", duration approx. 1:30 mins. CD recording: Callas-Gobbi/De Sabata 1953; video: Gheorghiu-Raimondi/NN).

A.2 Ensemble Scenes

- **Wolfgang Amadeus Mozart, Così fan tutte: Introduction, Recitativo Ferrando - Guglielmo - Don Alfonso**

Again, we start with a recitativo, taken from the introductory scene of Mozart's „Così fan tutte". Senior philosopher Don Alfonso (U) would like to convince his younger friends Ferrando (S) and Guglielmo (T) that their girlfriends are not faithful to them and proposes a bet on that („*Scioccherie di poeti!*", duration approx. 1:56 mins. CD recording: Araiza-Allen-van Dam/Marriner).

- **Wolfgang Amadeus Mozart, Magic Flute, 2nd Act: Trio Pamina - Tamino - Sarastro**

This examples is taken from Mozart's „Magic Flute". Princess Pamina (S) has to say good-bye to her boyfriend Tamino (U), while priest Sarastro (T) urges him to better be in time („*Soll ich dich, Teurer, nicht mehr seh'n?*", duration approx. 3:01 mins. CD recording: Lear-Wunderlich-Crass/Böhm 1964).

- **Charles Gounod, Faust, Act 5: Final Scene Marguerite - Faust - Mephistopheles**

The „Grand Opéra“ is represented by Gounod’s „Faust“. Having nearly managed to destroy Faust’s (U) relationship with Marguerite (T), Mephistopheles (S) tries to finish his plan but eventually fails due to divine forces („*Alerte, alerte!*“, duration approx. 3:37 mins. CD recording: Sutherland-Corelli-Ghiaurov/Bonynge).

- **Giuseppe Verdi, Rigoletto, 3rd Act: Quartet Gilda - Maddalena - Duca - Rigoletto**

One of the best-known ensembles in Italian opera is the quartet of Verdi’s „Rigoletto“. While the Duke of Mantova (S) seduces the prostitute Maddalena (U), the court jester Rigoletto (V) tries to convince young daughter Gilda (T) that the Duke is by no means worth her love („*Un dì, se ben rammentomi - Bella figlia dell’amore*“, duration approx. 5:13 mins. CD recording: Moffo-Elias-Kraus-Merrill/Solti).

- **Richard Strauss, Salome: „Judenquintett“**

The „Quintet of the Jews“ in Strauss’s „Salome“ is widely known as one of the most chaotic pieces in operatic literature. Motivated by a short argument between Jewish king Herodes (S) and his wife Herodias (T) about John the Baptist, five Jews (U, V, W, X, Y) and a Nazarenian (Z) intensively discuss his role as a prophet, before his voice brings them to silence („*Heiss ihn schweigen! Dieser Mensch beschimpft mich*“, duration approx. 3:38 mins. CD recordings: Kenney-Patzak-Braun et al./Krauss, and Hoffman-Stolze-Wächter/Solti 1962 for the user tests).

- **Giacomo Puccini, Gianni Schicchi: Introduction**

Our final example is taken from Puccini’s musical comedy „Gianni Schicchi“. The opera starts with the dead of rich Mr Donato, all of his relatives (from S to Z) being present and expressing their deep mourning while at the same time discussing how to proceed with his heritage. („*Ah! Ah! Povero Buoso*“, duration approx. 2:05 mins. CD recording: Alagna-Podles-Frittoli et al./Bartoletti).

REFERENCES

[1] G. Altmann: *Musikalische Formenlehre*. Saur Verlag Munich, 1981.

[2] P.T. Brady: *A Statistical Analysis of On-/Off-Patterns in 16 Conversations*. Bell Systems Technical Journal vol. 47 no. 1, pp. 73-91, Jan 1968.

[3] F. Hammer, P. Reichl, A. Raake: *Elements of Interactivity in Telephone Conversations.* Proc. ICSLP/INTERSPEECH’04, Vol. 3, pp. 1741-1744, Jeju Island, Korea, Oct 2004.

[4] F. Hammer, P. Reichl, A. Raake: *The Well-Tempered Conversation: Interactivity, Delay and Perceptual VoIP Quality.* Proc. IEEE Int. Conf. on Communications (ICC), Seoul, Korea, May 2005.

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

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

[7] S. Möller: *Assessment and Prediction of Speech Quality in Telecommunications.* Boston (Kluwer) 2000.

[8] P. Reichl, F. Hammer: *Hot Discussions and Frosty Dialogues: Towards a Temperature Metric for Conversational Interactivity.* Proc. ICSLP/INTERSPEECH’04, Vol. 3, pp. 1741-1744, Jeju Island, Korea, Oct 2004.

[9] P. Reichl, G. Kubin, F. Hammer: *A General Temperature Metric Framework for Conversational Interactivity.* Technical Report FTW-TR-2005-026, June 2005.

[10] P. Reichl, M. Balinova, F. Hammer: *Measuring Non-Spontaneous Interactivity - An Opera-related Case Study.* Accepted for: 5th Open Workshop of MUSICNETWORK - Integration of Music in Multimedia Applications. Vienna, Austria, July 2005.

[11] C. Roads: *The Computer Music Tutorial*. The MIT Press, 1996

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

[13] R. Strauss, C. Krauss: *Capriccio. Ein Konversationsstück für Musik in einem Aufzug.* Schott’s Söhne, Mainz, 1942.