International Telecommunication Union

# ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

# P.1305

(07/2016)

SERIES P: TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Telemeeting assessment

## Effect of delays on telemeeting quality

Recommendation ITU-T P.1305

ITU-T P-SERIES RECOMMENDATIONS

**TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS**

| | | |
|---|---|---|
| Vocabulary and effects of transmission parameters on customer opinion of transmission quality | Series | P.10 |
| Voice terminal characteristics | Series | P.30 |
| | | P.300 |
| Reference systems | Series | P.40 |
| Objective measuring apparatus | Series | P.50 |
| | | P.500 |
| Objective electro-acoustical measurements | Series | P.60 |
| Measurements related to speech loudness | Series | P.70 |
| Methods for objective and subjective assessment of speech quality | Series | P.80 |
| Methods for objective and subjective assessment of speech and video quality | Series | P.800 |
| Audiovisual quality in multimedia services | Series | P.900 |
| Transmission performance and QoS aspects of IP end-points | Series | P.1000 |
| Communications involving vehicles | Series | P.1100 |
| Models and tools for quality assessment of streamed media | Series | P.1200 |
| **Telemeeting assessment** | **Series** | **P.1300** |
| Statistical analysis, evaluation and reporting guidelines of quality measurements | Series | P.1400 |
| Methods for objective and subjective assessment of quality of services other than speech and video | Series | P.1500 |

*For further details, please refer to the list of ITU-T Recommendations.*

# Recommendation ITU-T P.1305

## Effect of delays on telemeeting quality


### Summary

Recommendation ITU-T P.1305 has the primary purpose of describing the impact of echo-free transmission delays on telemeeting quality of experience (QoE). Delay has a major impact on the interaction performance of conference participants, but is not always explicitly noticed by them; its impact being commonly attributed to other participants behaviour rather than being a system feature.

Before discussing the impact of the delays, it is useful to first consider how people behave when conversing in an ideal environment. Since this has been well studied in the linguistic discipline of conversation analysis (CA), a short discussion of conversations in that context and a discussion on the impact of delay from a CA perspective is included. Some analysis based on the alternative technique of conversation surface structures is also included.

A secondary purpose of this Recommendation is to provide guidance on appropriate testing methods for evaluating the effect of delay in telemeetings.



### History

| Edition | Recommendation | Approval | Study Group | Unique ID* |
|---|---|---|---|---|
| 1.0 | ITU-T P.1305 | 2016-07-29 | 12 | 11.1002/1000/12974 |

### Keywords

Conversation, conversation analysis, delay, subjective test, telemeeting.


* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

## FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

## NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

## INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

## Table of Contents

# Recommendation ITU-T P.1305

## Effect of delays on telemeeting quality

### 1 Scope

This Recommendation has the primary purpose of describing the impact of echo-free transmission delays on telemeeting quality of experience (QoE). Delay is special amongst communication impairments in that it can have a major impact on how well the users interact, but very often without them being explicitly aware of its presence. Typically users comment on its impact (e.g., "we kept on interrupting each other") or assume a delayed response is due to the other participants taking a long time to respond. Consequently its impact is not always captured in conversational-opinion tests. Crucially the impact has been shown to be highly dependent on the task or activity being undertaken.

The focus in this Recommendation is therefore on the human factor – its purpose is not to define specific quantitative limits or values, discuss the sources of delay or to discuss methods of delay measurement or methods to reduce its impact.

Before discussing the impact of the delays, it is useful to first consider how people behave when conversing in an ideal environment. Since this has been well studied in the linguistic discipline of conversation analysis (CA), a short discussion of conversations in that context and a discussion on the impact of delay from a CA perspective is included. Some analysis based on the alternative technique of conversation surface structures is also included.

A secondary purpose of this Recommendation is to provide guidance on appropriate testing methods. A number of these have already been described in Annex D of [ITU-T P.1301] and will not be repeated here. However, there have been a number of new developments since that recommendation came into force and it is appropriate to expand on that knowledge.

### 2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T G.114] Recommendation ITU-T G.114 (2003), *One-way transmission time*.

[ITU-T G.131] Recommendation ITU-T G.131 (2003), *Talker echo and its control*.

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

[ITU-T P.805] Recommendation ITU-T P.805 (2007), *Subjective evaluation of conversational quality*.

[ITU-T P.920] Recommendation ITU-T P.920 (2000), *Interactive test methods for audiovisual communications*.

[ITU-T P.1301] Recommendation ITU-T P.1301 (2012), *Subjective quality evaluation of audio and audiovisual multiparty telemeetings*.

[ITU-T P.1312] Recommendation ITU-T P.1312 (2016), *Method for the measurement of the communication effectiveness of multiparty telemeetings using task performance*.

[ITU-R BT.1359-1] Recommendation ITU-R BT.1359-1 (1998), *Relative timing of sound and vision for broadcasting*.

## 3 Definitions

### 3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

**3.1.1 telemeeting** [ITU-T P.1301]: A meeting in which participants are located at least two locations and the communication takes place via a telecommunication system. The term telemeeting is used to emphasize that a meeting is often more flexible and interactive than a conventional business teleconference and could also be a private meeting. The telemeeting could be audio-only, audiovisual, text-based, or a mix of these modes.

**3.1.2 conversational quality** [ITU-T P.1301]: The perceived quality when two or more test participants have a conversation.

**3.1.3 two-party** [ITU-T P.1301]: Two persons. Example: Two persons are participating in a telemeeting, having a conversation, performing a test task together, etc. If not explicitly stated differently, two-party implicates that the persons are at two locations.

**3.1.4 multiparty** [ITU-T P.1301]: More than two persons. Example: More than two persons are participating in a telemeeting, having a conversation, performing a test task together, etc. The term multiparty does not specify if the persons are distributed across two or more locations. If not explicitly stated differently, multiparty implicates that the persons are at two or more than two locations. When further specification is necessary, additional terms will be used (see point-to-point and multi-point) or the number of locations will be explicitly stated.

**3.1.5 point-to-point** [ITU-T P.1301]: Two locations. Example: A multiparty point-to-point telemeeting means that more than two interlocutors are taking part, and the interlocutors are at exactly two locations. That means that in one location more than one interlocutors are present, as there are more than two persons.

**3.1.6 multi-point** [ITU-T P.1301]: More than two locations. Example: A multiparty multi-point telemeeting means that more than two interlocutors are taking part, and the interlocutors are located across more than two locations. Multi-point does not specify if one or more than one interlocutor may be present at each location. In the special case that only one person is present at each location, the term one-per-site will be used.

**3.1.7 one-per-site** [ITU-T P.1301]: One person per connected location. Example: In a multiparty one-per-site telemeeting more than two sites are connected with only one person present at each site.

**3.1.8 conversation analysis (CA)** [b-Sacks, 1974]: The study of social interaction, in particular focusing on conversations.

**3.1.9 turn construction unit (TCU)** [b-Sacks, 1974]: A conversation analysis term describing the fundamental segment of speech in a conversation – essentially a piece of speech that constitutes an entire 'turn'.

**3.1.10 transition relevance place (TRP)** [b-Sacks, 1974]: A conversation analysis term which indicates where a turn or floor exchange can take place between speakers.

**3.1.11 system effectiveness** [ITU-T P.1312]: The ratio, expressed as a percentage, between the performance score achieved on a given system and the performance score obtained in face-to-face communication.

### 3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

**3.2.1 choral behaviour**: Many people doing the same thing at the same time.

**3.2.2 phatic expression**: An expression whose function is part of social interaction rather than to convey information.

## 4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

CA Conversation Analysis

MOS Mean Opinion Score

QoE Quality of Experience

QoS Quality of Service

TCU Turn Constructional Unit

TRP Transition Relevance Place

## 5 Conventions

None.

## 6 Existing Recommendations concerning the subjective quality evaluation in case of pure transmission delay

In the following existing Recommendations on audio delay, video delay and audio-video synchronization the current understanding is outlined. These Recommendations have been developed for two-party interaction only. The multi-party case needs further study to ensure their applicability.

### 6.1 Audio delay

Recommendation [ITU-T G.114] states that,

"Regardless of the type of application, it is recommended to not exceed a one-way delay of 400 ms for general network planning…"

"Although a few applications may be slightly affected by end-to-end (i.e., "mouth-to-ear" in the case of speech) delays of less than 150 ms, if delays can be kept below this figure, most applications, both speech and non-speech, will experience essentially transparent interactivity."

The delay values in [ITU-T G.114] are based on the effect of pure audio delay only, i.e., in the complete absence of any echo. It is well-known that if echo is present, longer audio delays become easier detectable and much more disturbing. Effects of talker echo at different one-way echo-path delays are further described in [ITU-T G.131].

### 6.2 Video delay

Most investigations regarding quality impact of delay have been performed regarding audio delays, but in an earlier version of [ITU-T G.114] (02/96), in clause B.2.3, it is mentioned that:

"Tests were performed to assess the interaction between delay and user applications. In these tests a comparison of telephone conversations with videophone were made and it was shown that there is little difference between both types of connection."

This result was obtained for synchronized audio and video. Some research has been performed looking at pure audio-visual delay in an interactive setting, but there is a need for more research to clearly confirm the recommendations of [ITU-T G.114] for this case, especially regarding multi-party communication.

### 6.3 Synchronization between audio and video

The limits within which audio and video are perceived to be synchronized are well examined, at least for TV screens. According to [ITU-R BT.1359-1],

"subjective evaluations show that detectability thresholds are about +45 ms to −125 ms and acceptability thresholds are about +90 ms to −185 ms on the average, a positive value indicates that sound is advanced with respect to vision".

As video encoding usually requires more time than audio coding, audio should generally be delayed to obtain audio-video synchronization [b-Berndtsson, 2012]. It is likely that it is easier to identify audio-video asynchrony as a technical issue than synchronous delay. Following from this, audio-video asynchrony will most likely be reflected more critically in mean opinion score (MOS) ratings.

Low video frame rates and video frame rate jitter can influence the impression of audio-video synchronization.

## 7 Conversation features

### 7.1 Two-party and multi-party conversations

There are differences between two-way conversations and multi-party conversations. For instance in multi-party conversations:

– participants are likely to spend a greater proportion of the time listening rather than talking. Usually this will vary as the discussion proceeds, with talkers being more active or passive at different times;

– telemeeting systems do not always attempt to reproduce the exact signals from the sending end at the receiving end; system designers are usually more interested in creating an environment where participants can interact effectively, and this can involve multiple technologies such as spatial audio, level compensation, noise reduction etc.;

– participant interaction is more complicated: in a two-party conversation the role of the speaker and the listener simply alternate, whereas in a multi-party conversation, the next speaker could be any one of several current listeners;

– passive participants are not only listening to what the active participants are saying, they are listening to the active participants interacting.

### 7.2 Two analysis techniques

In this Recommendation two different techniques for analysing a conversation are discussed. The first is based on the linguistic discipline of conversation analysis (CA) [b-Sacks, 1974] and focuses on analysing how the participants interact. For evaluating the impact of delay, the two most relevant analysis concepts from CA are turn-taking (see clause 7.3) and backchannels (see clause 7.6).

The second technique analyses patterns of the temporal occurrence of utterances without labelling their purpose. This approach does not rely on the content of the utterances and can therefore be

built on voice activity detection algorithms. This approach is referred to as conversational surface structure analysis (clause 7.4) and was pioneered by Brady in the 1960s. [b-Brady, 1965].

### 7.3 Conversation analysis and turn-taking

In a free conversation the organization of the conversation, in terms of who speaks when, is referred to as 'turn-taking'. This is implicitly negotiated by a multitude of verbal cues within the conversation and also by nonverbal cues such as physical motion and eye contact. Turn-taking behaviour is more-or-less universal, even across languages of very different structures. This behaviour has been extensively studied in the discipline of conversation analysis, for example, [b-Sacks, 1974].

Some useful concepts from CA are:

– The turn constructional unit (TCU), which is the fundamental segment of speech in a conversation – essentially a piece of speech that constitutes an entire 'turn'. TCUs are primarily determined as being pragmatically or grammatically complete utterances. That is, have they served a specific purpose, or are they syntactically complete. Typically they are separated by short pauses and a falling tone.

– The transition relevance place (TRP), which indicates where a turn or floor exchange can take place between speakers. TCUs are separated by TRPs.

– Turn allocation component, which relates to who should speak next. This is done either explicitly when the current talker selects the next talker – directly or subtlety; or the next talker 'self-selects' by simply interrupting at the end of the TCU.

These processes enable the basic turn-taking process to take place, as described in [b-Sacks, 1974] and shown in a slightly modified form in Figure 1. As a TCU comes to an end the first decision to take place is, has the current talker selected the next talker? If so, the designated new talker usually feels obliged to talk. Otherwise, any participant including the current talker will self-select; a process which generally works by the first person to speak gaining the right to the next turn. The result of this process is a well ordered conversation as illustrated in Figure 2.

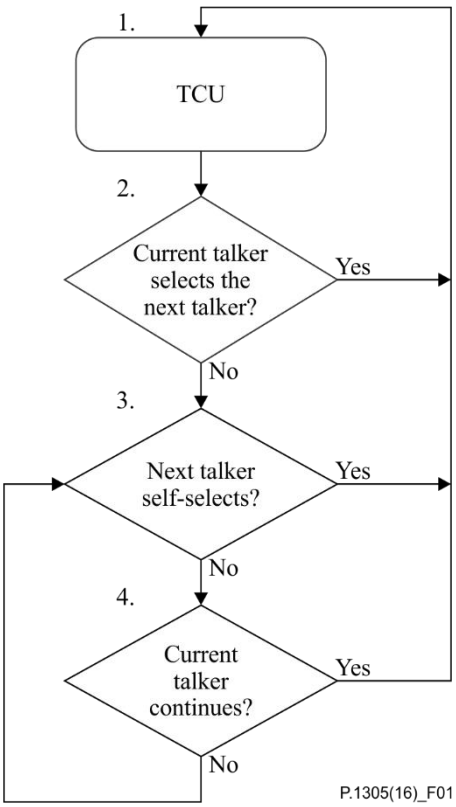


**Figure 1 – The conversation turn-taking process**

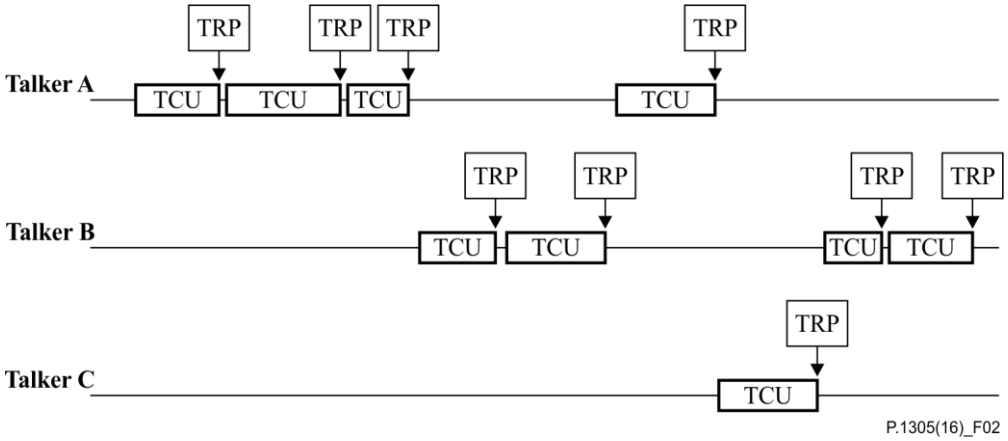


**Figure 2 – Example of turn-taking in a well-ordered conversation**

Note of course that most of the turn-taking process is carried out subconsciously in line with what the participants consider to be appropriate behaviour. Knowing when it is acceptable to take your turn in a conversation is usually learned from an early age and becomes an important social skill.

Depending on the type of conversation, there may also be a significant amount of talking taking place outside the underlying turn-taking process – in some cases obscuring it completely. This can include backchannels (clause 7.5), brief side-conversations, corrections, confirmations and choral laughter.

Note that this description of the turn-taking process excludes interruptive behaviour. From a CA perspective, a participant starting a turn at a time other than a TRP is making a 'violative interruption'. This can often be interpreted by other participants as ill-mannered or rude behaviour.

This turn-taking decision process leads to the following meeting characteristics:

– Overwhelmingly, only one participant talks at a time [b-Sacks, 1974].
– Occurrences of more than one talker at a time are common, but brief [b-Sacks, 1974].
– Transitions from one turn to the next, with no gap or overlap are common. Together with transitions characterized by either a slight gap or slight overlap, they make up the vast majority of transitions [b-Sacks, 1974].
– The most frequent gaps between talkers are in the region of 200 ms. [b-Gisladottir, 2015]. Gaps of more than 1 second are rare but do occur. [b-Jefferson, 1989]

### 7.4 Conversation structure and styles

Conversations can vary dramatically in their style of structure. At one extreme relaxed free conversation can occur amongst friends in social situations, where all turn taking is organized as described above. There may also be a significant amount of talking taking place outside the underlying turn-taking process, in some cases obscuring it completely. This additional content can include backchannels (clause 7.5), brief side-conversations, corrections, confirmations and choral laughter.

At the other extreme some conversations are highly structured where most turn-taking is explicitly directed by a chairman or moderator. In this case a participant is directly asked by the participant currently holding the turn to speak.

A conversation can change style dynamically, e.g., following an agenda but breaking out to free conversations in the discussion of single items.

### 7.5 Backchannels and other behaviours during conversations

A further aspect to consider is that there is often a lot of talking taking place outside the formal turn-taking process as described in CA. A conversation feature of particular interest is the use of continuers or backchannels [b-OConnail, 1993], also referred to as 'continuers' or' listener responses' [b-Knapp, 2010]. These are listener responses in a primarily one-way communication, commonly used by participants in a conversation as a sign of acknowledgement and also to signal to the talker to continue with their turn. These can be both verbal and non-verbal in nature (e.g., "*yeah*", " *mm*", " *uh-huh*", " *mm-hmm*") and are frequently phatic expressions. They are often made between sentences and especially at transition points. Backchannels can also be visual in the form of nods or other facial expressions.

The term "backchannel" was designed to imply that there are two channels of communication operating simultaneously during a conversation; the predominant channel is that of the speaker who directs primary speech flow.

Other behaviours include: requests for confirmation or clarification; choral behaviour, especially laughter; and side conversations between subgroups. In many situations, particularly informal ones, it is not unusual to have several conversations taking place simultaneously with participants both talking and listening at the same time. Thus the ability of a telemeeting system to support doubletalk is a very important feature for natural human communications.

### 7.6 Conversation surface structure analysis

Conversation surface structure analysis has been introduced by Brady [b-Brady, 1965], [b-Brady, 1968], [b-Brady, 1971]. It analyses the temporal occurrence of speech activity, that is, on-off patterns of speech and uses this to define various conversation states such as single-talk, multiple-talk. These are then analysed to provide metrics such as:

– State sojourn time: the mean time of staying in that state, [b-Brady1971]
– Speaker alteration rate: the mean rate at which speaker changes occurred [b-Hammer, 2004])

– Utterance Rhythm: mean time from one utterance of a speaker to its next utterance, [b-Schoenenberg, 2014].

Conversation surface structure analysis consists of the following procedure:

1. A voice activity algorithm can be used to extract those on-off patterns of each single person recording. See for example Figure 3, top panel.
2. Then for each site, the single person on-off patterns need to be combined as they were experienced by the persons at the different ends which means the respective other persons need to be delayed if delay was present. See for example Figure 3, bottom panel.
3. In the next step, states need to be assigned to every sample of the combined on-off pattern, according to Figure 4. In this case 'S' and 'M' are used to indicate single and multiple talker states, and sub-scripts 'n,X' to indicate how many talkers there are and which one is talking. Thus for two participants. $S_{2,0}$: silence, $S_{2,A}$: single talk of person A, $S_{2,B}$: single talk of person B, $M_{2,AB}$: multi/double talk of persons A and B. State transitions are indicated by arrows.
4. From those states, it is possible to compute the measures such as state sojourn time, speaker alteration rate and utterance rhythm that characterize the conversation surface structure.

An example of how to do this is given in clause 10.

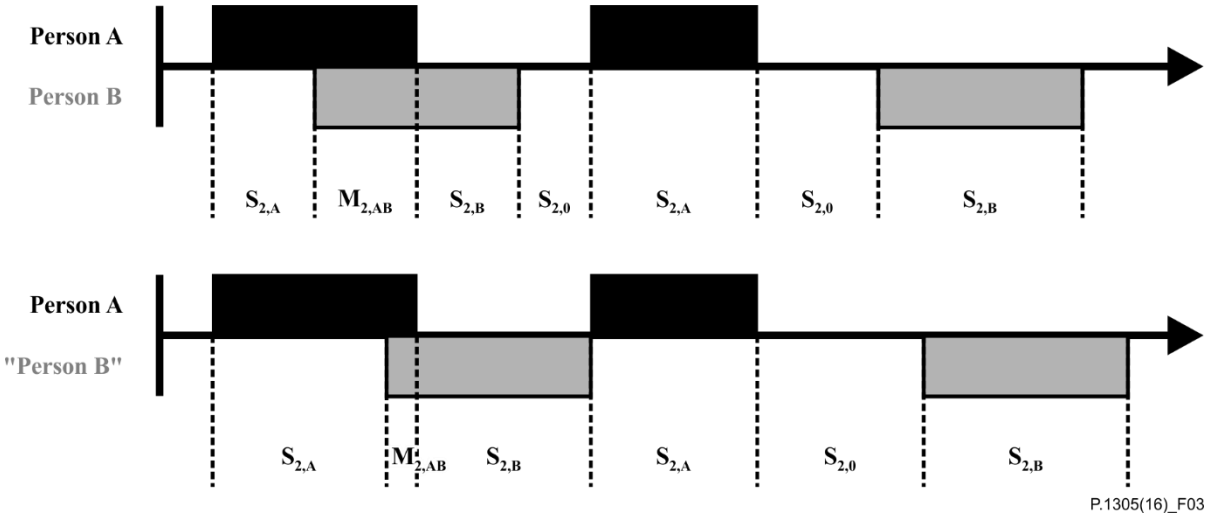


**Figure 3 – Example of single and multi talk states in a two-party conversation situation. Top panel: Situation without delay. Bottom panel: Situation from the perspective of Person A when delay is present, i.e., person B is delayed**

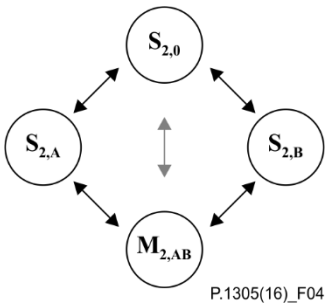


**Figure 4 – Possible states (sets and subsets) for a two-party conversation situation**

# 8 The impact of delay

## 8.1 Turn-taking with delay

From a conversation analysis perspective, the impact of delay in a conversation is primarily to disrupt the turn-taking process because each participant has a different understanding of when the TRP occurs. A new talker may believe they are starting to talk at the TRP, but the other participants will hear them sometime later when they themselves may have already started talking. This breaks the self-selection process described in clause 7.3 because the participants cannot be aware of who started talking first. Essentially, none of the participants is able to interrupt at the TRP experienced by the other participants, and if the delay is long both talkers interpret the others action as a violative interruption.

Note that the participants may not realize that the confusion they are experiencing is due to transmission delay, often attributing it to aberrant behaviour of the other participants, for example, slow responding or even rudeness.

A good example of this delay induced confusion is the 'false start'. These are delay induced conversation features that occur when two participants start talking simultaneously after a period of silence or in response to a stimulus from another participant. For example, in a multi-party conversation this can occur in response to a question from a third participant as shown in Figure 5. Once they realize somebody else is also talking, one of three things can happen:

1. Both participants stop talking. This could be interpreted as two simultaneous unintended successful interrupts. More false starts can then follow in an attempt to correct the false-start, possibly involving other participants, leading to a short period of confusion as to who actually has the conference 'floor'. From a CA point of view this has a major impact since the conversation stops unexpectedly.
2. One participant stops talking and the other continues. Whilst this is probably the least disruptive on conversation flow, it can still be a significant disruption of the information exchange; the participant who stopped talking may have been making a more useful contribution than the participant who continued. It has also provided a significant distraction from the matter under discussion.
3. Both participants keep talking resulting in a longer period of multi-talk. Whilst the conversation still continues, the impact on intelligibility is severe and from a social point of view the behaviour of the participants can be perceived as rude or inappropriate.

False starts are common in multi-party telemeetings where delay is present. However normal conversation turn-taking makes them less likely in two-way conversations. Of course two participants can start talking simultaneously even if there is no delay. However, delay appears to make the problem worse because it takes longer for the situation to become apparent to the participant.

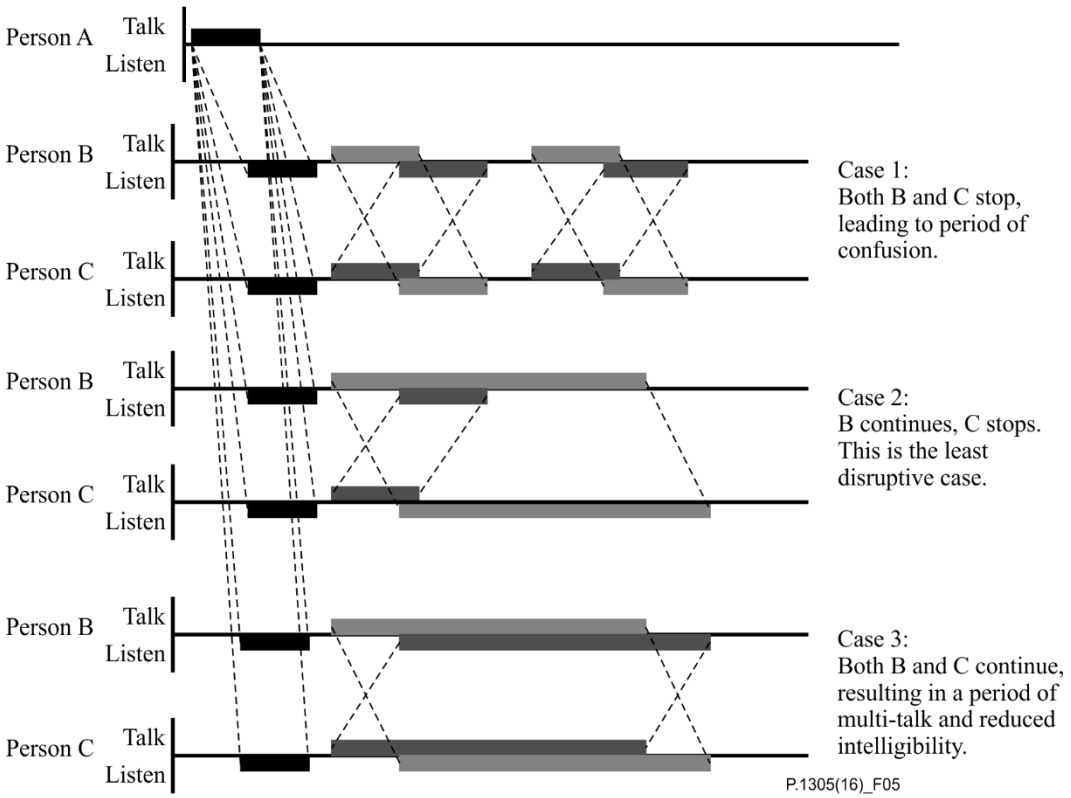


**Figure 5 – False starts in 3-way conversation**

## 8.2 Conversation surface structure and delays

It has been shown in many studies e.g., [b-Brady, 1971], [b-Egger, 2012], that the conversation flow is disrupted by delays. In particular, it has been described in different works [b-Brady, 1971, [b-Schoenenberg, 2014] that the quality of the conversational experience compared to a face-to-face conversation diverges in terms of conversational realities of each telemeeting participant, and that the degree of divergence is dependent on the length of the delay [b-Schoenenberg, 2014]. The higher the degree of divergence, the more difficult it will be for participants to hold a fluent interaction without misunderstandings or breakdowns of the interactional flow.

An example of a conversation part with high divergence can be found in Figure 6.

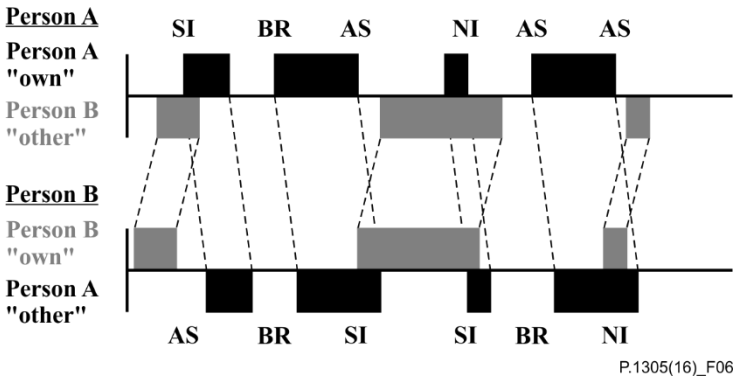


**Figure 6 – Example of conversation part with high degree of divergence between the two realities due to latency**

Here it can be seen that owing to the delay the number of states has increased from the four states defined in clause 7.6, to include breaks (BR), alternating silence (AS), successful interruption (SI)

non-successful interruption (NI). An explanation of how the degree of divergence figure can be calculated is given in clause 10.3.

### 8.3 Multiparty and two-party conversations with delay

So far there is only little research on multiparty conversations and delay; most research conducted has focused on the two-party case. However, it is likely the impact of delays will be different in multi-party conversations for the following reasons:

– Participants are likely to spend a greater proportion of the time listening rather than talking. This will vary as the discussion proceeds, with talkers being more active or passive.

– Active participants, those actively contributing to the discussion, have shown to be more sensitive to delay compared to their non-active fellow participants [b-Schmitt, 2014a], [b-Schoenenberg, 2011].

– The turn-taking process is more complicated. For example, in a two-party conversation the role of the speaker and the addressee alternate. In multi-party conversations, the speaker can have multiple addresses and listeners and thus more responders. A greater likelihood of false starts or passive interruptions can thus be expected.

### 8.4 Other side-impacts of delay

Delay can also impact the performance of the system, for example, it can affect jitter buffers and result in packet loss which in turn can lead to audible artifacts.

## 9 Factors influencing the impact of pure delay

In clause 7, it was noted that conversations can differ greatly in their structure, ranging from formal meetings with strict control, up to very informal social gatherings, and that the impact of delay will differ in each case. In this section other factors that influence the impact of delay are considered.

### 9.1 Audiovisual telemeetings

The previous clauses have primarily discussed the audio-only meetings. In the case of synchronized audiovisual telemeetings there are the following differences:

– Delay is more obvious to non-naive participants since visual reactions are seen.

– Visual cues (e.g., gestures, nods, gaze) are available to support turn-taking and back-channel communications.

– The video provides additional information that helps the conversation organization. For example, participants can make visual signals to each other: e.g., raising hand to interrupt, collective voting.

Subjective tests have shown that it is more important to synchronize audio and video than to keep the one-way audio transmission delay as low as possible [b-Berndtsson, 2012]. Since video is usually more delayed than audio, the audio should generally be delayed to obtain audio-video synchronization, at least for transmission delays below 600 ms. The audio should not be more than 90 ms ahead of the video, and also less than 185 ms after the video. This reflects the natural experience of light travelling faster than sound [ITU-R BT.1359-1].

A study [b-Tam, 2012] comparing two-party conversations with audio and audiovisual channels, found that perceived conversation naturalness was significantly impacted at a delay of 400 ms for audio-only connections and at 500 ms for audio-visual connections. Similar insights were drawn from studies comparing face-to-face, audio-only and audiovisual meeting [b-Tang, 1992]. In audiovisual meetings longer silences were observed but the subjective scores from the participants did not reflect this.

## 9.2 User expectations

Whilst user expectations will have little impact on the effect of delay in a telemeeting, users are likely to be more accepting of delays under some conditions. For example they will have lower expectations from lower cost services or when making calls on more challenging situations.

The physical distance also adds to the delay and users are usually aware of this. For example, signals propagate through physical media at approximately 2/3 of the speed of light, so it takes about 50 ms to travel 10,000 km. Therefore expectations might be lower when talking to someone on the other side of the world than when talking to someone in the office next door.

## 9.3 Asymmetric delay between participants and co-location

Usually, there will be different delays in the connections to each site. In this situation, test participants experiencing longer delays will have more difficulties taking part in the conversation. This can have the knock-on effect of reducing the quality of the whole group experience. It can, for example, cause a disruption of the conversation flow, or as it has been shown [b-Schmitt, 2014b], a single delayed person can cause a worse experience for all participants in a five party telemeeting.

A similar situation will occur if many participants are co-located and only a few are joining over a delayed connection.

## 9.4 Interpersonal influences

The sensitivity for delay can be influenced by the participant's expectations of the other participants. It can be difficult for participants to determine whether an apparently slow response is due to the system or to a genuine slow response. To a certain extent the degree of familiarity with the other participant can influence this based on previous interactions with them.

However, it has been shown that conversation partners that know each other can be perceived to be unusually slow if there are long delays [b-Berndtsson, 2012]. Furthermore, persons that know each other well often have more spontaneous conversations, which can also make delays more detectable. Note that even though interruption may be a part of the normal conversations in this case, they do not occur with awkward timing or too often.

The social role of the interlocutors and their corresponding interaction behaviour can also have an impact. For instance, a dominant person, the boss or mentor, may speak more and may ask more questions within a call. If this is the case, response times will appear to be long for the asking person but short for the responding person, which leads to different perceptions for the different sites.

# 10 Testing methods

## 10.1 General comments

It is recommended to carry out the subjective quality evaluation of multiparty telemeetings as much as possible according to existing test methods recommended by ITU-T.

Since the primary impact of delay is on group interaction, it will be necessary to use some form of conversation in the test. This could include multi-party conversation tests or third party listening tests where subjects listen and comment on the quality of a conversation between others.

Tests on group inter-action can be divided into several types:

– conversational-opinion tests;

– group inter-action tests based on conversation analysis and conversation surface structure analysis;

– tests that measure group performance when undertaking tasks;
– micro-feature tests such as the task efficiency tests described in [ITU-T P.1312] which focus on a specific aspect of talker interaction such as double-talk or turn-taking.

It is possible to design tests that include elements of more than one of these types.

## 10.2 Conversational-opinion tests

General procedures for these are laid down in Annex D of [ITU-T P.1301]. The test must allow the subjects to experience the problem and report its impact in terms of conversation difficulty.

### 10.2.1 Training session

In most cases, the test procedure will include a short training phase to familiarize the subjects with the system and the test task. This is important if the subjects do not know each other well in order to encourage more interactive conversations. This could be incorporated into an introduction session if appropriate and consist of a simple interactive game or task to learn each other names and to generally 'break the ice'.

### 10.2.2 Conversational test tasks

The test tasks should ideally be delay sensitive, but support natural conversation. Some suggestions for appropriate tasks are given in Annex D of [ITU-T P.1301] and in Annex A. More specifically:

– Free conversation tasks are recommended for studying group behaviour since they allow natural subject interaction and subjects are able to adapt their behaviour to make the best use of the system under test.
– Short conversation tests (SCTs) as described in [ITU-T P.805] are not delay sensitive because subjects have to read from a script which masks the delay [b-Berndtsson, 2012].
– Highly interactive tasks such as random number exchanges are sensitive to delay, but do not reflect normal conversation behaviour.
– Delays can be more easily detected as a technical issue in a highly interactive conversation where the motivation to interact fast is high [b-Schoenenberg, 2014]. As an example, some kind of competition might motivate subjects to interact more efficiently, and delays may be more apparent if they are hindering the interaction.

### 10.2.3 Test questions

Care should be used in the choice of questions in order to enable the test subjects to appropriately comment on their experience. For example, questions based on how the subjects perceived the interactivity are likely to be more revealing than questions about the overall quality.

For example:

– Did you experience difficulty in communicating?
– How would you judge the effort required to interrupt the other participants?
– How easy did you find it to communicate using the system?
– How attentive were your conversation partners during the call?

A number of appropriate examples are given in [ITU-T P.805] and Annex D of [ITU-T P.1301], along with accompanying scales.

Because delay induced interrupts are sometimes perceived as rude it would be good if these perceptions could also be captured. This can be difficult since subjects may object to their individual personalities being scrutinised in this manner. Perhaps the best approach, if the subjects knew each other well, is to ask questions such as, "Did the behaviour of any of the other participants seem out of character at any time?

## 10.3 Group interaction tests

The principle behind group interaction tests lies in a comparison between the behaviour of test subjects in a reference condition, with their behaviour when using the system under test.

Of the two approaches discussed earlier, conversation analysis, and conversation surface structure analysis, the latter is more mature and its focus on simply who is talking at any given time, lends itself more to an algorithmic approach [b-Schoenenberg, 2014]. Because conversation analysis assumes an understanding of the actual conversation content, detailed analysis is likely to require higher level resources such as speech recognition and natural language processing. Consequently at this stage no measurement techniques directly based on it have been published. However, analysis based on simple turn-taking could be considered.

In group interaction tests, in addition to normal good practice in subject selection [ITU-T P.1301], [ITU-T P.800], careful consideration should be given to matching test subject profiles. This is important for free conversations since as far as possible we need to ensure consistent behaviour across groups.

The number of variables in free conversations that influence conversational interaction can be very large. The interactivity in the conversation might be higher or more consistent if the test subjects have:

– The same first language.
– A similar level of experience in using telemeeting systems.
– Similar levels of familiarity with each other.
– Similar cultural backgrounds.

It may also be appropriate to consider matching ages, but of course, note that excessive restrictions placed on subject selection will reduce the generality of the results.

### 10.3.1 Conversation surface structure measurements

It can be seen from clause 8.2 that the quality of the conversational experience compared to a face to face conversation diverges in terms of conversational realities of each telemeeting participant, and that the degree of divergence is dependent on the length of the delay [b-Schoenenberg, 2014]. A measure of this divergence can be calculated using the following procedure, and with reference to Figure 6:

1. First of all, the most used meaningful state walks needs to be defined, that are:

   BR: Break – After a single talk period of one person, a silent period follows, after which single talk of the same person continues.

   AS: Alternating silence – After a single talk period of one person, a silence follows, after which single talk of a different person continues.

   NI: Non-successful interruption – After a single talk period of one person, double talk or multi talk (e.g., triple talk) follows, after which single talk of the same person continues.

   SI: Successful interruption – After a single talk period of one person, double talk or multi talk (e.g., triple talk) follows, after which single talk of a different person who has started to talk in the double/multi talk period continues.

2. Compute for each person a speaker state pattern using the conversation surface structure analysis paradigm (clause 7.4) as it was experienced by the persons at the different ends which means the respective other persons need to be delayed if delay was present. See Figure 6.

3. Based on the state pattern, state walks can be detected. For each walk detected in i.e., the conversational reality of person A it is then looked into the same walk at the reality of person B to see how person B experience this walk. The origin for each walk is always a

single talk period. When looking at person A's reality e.g. single talk of person A, followed by a silence, followed by single talk of person A, this will be noted as a break by person A in A's reality. In case of two interlocutors, eight different kinds of walks can occur, the four types listed above, starting with either person A or person B in single talk:

| State walk | Description | States transitions (see Figure 4) | Speaker turn |
|---|---|---|---|
| BR.AA | Break, starting and ending with A | $S_{2,A} \Rightarrow S_{2,0} \Rightarrow S_{2,A}$ | No |
| BR.BB | Break, starting and ending with B | $S_{2,B} \Rightarrow S_{2,0} \Rightarrow S_{2,B}$ | No |
| AS.AB | Alternating silence, starting with A, ending with B | $S_{2,A} \Rightarrow S_{2,0} \Rightarrow S_{2,B}$ | Yes |
| AS.BA | Alternating silence, starting with B, ending with A | $S_{2,B} \Rightarrow S_{2,0} \Rightarrow S_{2,A}$ | Yes |
| NI.AA | Non-successful interruption, starting and ending with A | $S_{2,A} \Rightarrow M_{2,AB} \Rightarrow S_{2,A}$ | No |
| NI.BB | Non-successful interruption, starting and ending with B | $S_{2,B} \Rightarrow M_{2,AB} \Rightarrow S_{2,B}$ | No |
| SI.AB | Successful interruption, starting with A, ending with B | $S_{2,A} \Rightarrow M_{2,AB} \Rightarrow S_{2,B}$ | Yes |
| SI.BA | Successful interruption, starting with B, ending with A | $S_{2,B} \Rightarrow M_{2,AB} \Rightarrow S_{2,A}$ | Yes |

Similarly, such state walks can be computed for multiparty scenarios. Here examples for a three-party call for state walks starting with person A:

| State walk | Description | States transitions (see Figure 4) | Speaker turn |
|---|---|---|---|
| BR.AA | Break | $S_{3,A} \Rightarrow S_{3,0} \Rightarrow S_{3,A}$ | No |
| AS.AB | Alternating silence | $S_{3,A} \Rightarrow S_{3,0} \Rightarrow S_{3,B}$ | Yes |
| AS.AC | Alternating silence | $S_{3,A} \Rightarrow S_{3,0} \Rightarrow S_{3,C}$ | Yes |
| NI.AA | Non-successful interruption | $S_{3,A} \Rightarrow M_{3,AB}\ (\Leftrightarrow M_{3,ABC} \Leftrightarrow) \Rightarrow S_{3,A}$ or: $S_{3,A} \Rightarrow M_{3,AC}\ (\Leftrightarrow M_{3,ABC} \Leftrightarrow) \Rightarrow S_{3,A}$ | No |
| SI.AB | Successful interruption | $S_{3,A} \Rightarrow M_{3,AB}\ (\Leftrightarrow M_{3,ABC} \Leftrightarrow) \Rightarrow S_{3,B}$ or: $S_{3,A} \Rightarrow M_{3,AC}\ (\Leftrightarrow M_{3,ABC} \Leftrightarrow) \Rightarrow S_{3,B}$ | Yes |
| SI.AC | Successful interruption | $S_{3,A} \Rightarrow M_{3,AB}\ (\Leftrightarrow M_{3,ABC} \Leftrightarrow) \Rightarrow S_{3,C}$ or: $S_{3,A} \Rightarrow M_{3,AC}\ (\Leftrightarrow M_{3,ABC} \Leftrightarrow) \Rightarrow S_{3,C}$ | Yes |

In the next step the first period of single talk of a detected walk in one person's reality is searched in the other persons' reality and the walk that is found there is noted. It can be the same kind of walk or a different one. In this way, the number of matching walks and mismatching walks can be counted. A convenient approach for this is to use a confusion matrix as shown below: the rows represent the walks found in the "own" reality of a person, e.g., person A; the columns represent the walks found in the reality of the other person, e.g., person B; and the matrix elements contain the counts. In such a matrix, the diagonal represents the matching walks (indicated with •), all other matrix elements represent the non-matching walks (indicated with ×).

| | | Reality of person B | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **BR.AA** | **BR.BB** | **AS.AB** | **AS.BA** | **NI.AA** | **NI.BB** | **SI.AB** | **SI.BA** |
| Reality of person A | BR.AA | ● | × | × | × | × | × | × | × |
| | BR.BB | × | ● | × | × | × | × | × | × |
| | AS.AB | × | × | ● | × | × | × | × | × |
| | AS.BA | × | × | × | ● | × | × | × | × |
| | NI.AA | × | × | × | × | ● | × | × | × |
| | NI.BB | × | × | × | × | × | ● | × | × |
| | SI.AB | × | × | × | × | × | × | ● | × |
| | SI.BA | × | × | × | × | × | × | × | ● |

In case of a multiparty scenario, this confusion matrix is extended with the additional state walks and it is computed for each pair of the reality of two persons.

Then, from each such confusion matrix between two persons X and Y, the total number of matching and non-matching walks can be computed: numMatch_XY = sum of counts in the diagonal elements numMismatch_XY = sum of counts in all non-diagonal elements

Finally, the number of mismatching walks needs to be related to the number of matching walks for building the parameter *divergence* [b-Schoenenberg, 2014], which is computed for each person using the corresponding confusion matrices.

For the two-party scenario this leads to Divergence_A = Divergence_B = numMismatch_AB / numMatch_AB

For the three-party scenario this leads to:
Divergence_A = (numMismatch_AB + numMismatch_AC) / (numMatch_AB + numMatch_AC)
Divergence_B = (numMismatch_AB + numMismatch_BC) / (numMatch_AB + numMatch_BC)
Divergence_C = (numMismatch_AC + numMismatch_BC) / (numMatch_AC + numMatch_BC)
The divergence is small if conversational realities are similar and high if they are rather different.

## 10.4 Communication system effectiveness tests

System effectiveness refers to the time that is needed to achieve the goals of a conversation using a communication system, relative to the time taken in a face to face meeting. Transmission delays will reduce the system effectiveness, increasing the meeting time not just because of the increased delay, but also because of the impact of the delay on the conversation turn-taking described above.

## 10.5 Micro-feature tests

An example of a micro-efficiency test is the task efficiency test [ITU-T P.1312] which is specifically designed to measure how well a system supports discussions with high levels of multi-talk. Whilst not designed to measure delay performance, it is delay sensitive since good performance in multi-talk relies on good interactivity behaviour which is in turn influenced by delay. This is shown in Appendix 3 of [ITU-T P.1312].

Other tests designed to focus on conversation features such as false starts or turn-taking can also be envisaged.

## 11 Areas for further research regarding the impact of delay

There are many aspects of the impact of delay not yet fully understood and warrant further study. These include:

1) The effects of asymmetric delay conditions between different participants in a telemeeting.
2) The effects of varying delays in a telemeeting.
3) Co-dependencies to other impairments such as echo, low frame rate, frame size.
4) The relationship between the impact of delay and the features of different types of audio rendering.
5) The impact of delay in relation to factors like the types of clients (screen size) and the number of participants using each client.
6) The relationship between delay and spatial audio rendering.
7) Analysis algorithms based on conversation analysis and/or turn-taking.

# Annex A

## Suggestions for free-conversation tasks

(This annex forms an integral part of this Recommendation.)

In [b-Berndtsson, 2012] it was noted that the best type of task for evaluating the impact of delay was a free conversation, where the subjects are free to express themselves in a natural manner, such as they would in normal life. However, it is also important to stimulate conversation of an appropriate style for the test. Typical conversation styles might include:

a. Relaxed, informal conversation.

b. Formal, 'business' style conversation.

c. Focused, problem solving discussion.

A range of test tasks for audio conversational tests are described in [ITU-T P.800] and [ITU-T P.805], with further test tasks for audiovisual conversational tests are described in [ITU-T P.920]. However, these are primarily tasks suitable for two-way conversations and are not all suitable for extension to the multi-party situation. A range of tasks more appropriate for multi-party tests is given in Appendices III and V of [ITU-T P.1301].

### A.1 'Holiday' task

In this task the participants are asked to:

a. Create a list of features that their ideal holiday would include (5 minutes).

b. List them in order of importance (5 minutes).

This task approach can be extended to include features of an ideal job, pastime, film, etc.

Good points: A very good task for getting people involved in an informal type meeting – people like talking about experiences they enjoy.

Poor points: It does not guarantee a balanced conversation – some participants could dominate or be too passive.

### A.2 Role-playing games

In this the test subjects take part in a role in a fictitious scenario. This is not a scripted test such as SCT, but relies on the test subjects taking on a role in a scenario. A good example of this is the "You be the Judge" game, (http://www.amazon.co.uk/Paul-Lamond-You-Be-Judge/dp/B00422MKG4) where the test subjects take on roles in a courtroom such as judge, council for the prosecution, defence, defendant, etc., and role play a series of scenarios in order to reach a solution.

Good points: Stylistically this tends to replicate business type formal meetings.

Poor points: Relies on subjects acting abilities and willingness to play a sufficiently active part.

### A.3 "Who am I?" Celebrity Name-Guessing Task

This task has been proposed in [b-Schoenenberg, 2014b] as a task for audiovisual communication with high delay sensitivity. In this task, participants are supposed to be a celebrity, and the task is to guess their own name.

The task should not be confused with the *Name-Guessing Task* described in [ITU-T P.920]. The *Celebrity Name-Guessing Task* described here has strong social component due to the aspect that a participant is associated with a celebrity that he or she has to guess. In addition, the game character of this task let people find themselves in a joyful state, have fun and interact naturally.

The game rules are described in [b-Schoenenberg, 2016] as follows:

> *Each participant receives a note with initials of a celebrity person that he or she has to guess. This note is clipped to the shirt of the person so that they and their interlocutor are reminded of it. The respective other knows the full name of the unknown celebrity person due to a hidden note with the full name. One participant begins by asking questions that can only be answered with "yes" or "no", for example, "Am I female?" or "Am I a movie star?". As long as the response of the other one is "yes" the participant is allowed to keep asking. If the answer is "no" the roles of asking and answering person are reversed. In this way, the roles are switched several times until one person correctly guesses the name. If one participant wins a game he or she gets a point. After the entire session (nine games) the participant who won most games wins the entire session and with this an additional prize (a voucher).*

Advantage(s): Let people use both audio and video channel to communicate. In particular, the primarily social interaction in this task facilitates visual communication in terms of nonverbal signals such as smiles (and generally facial expressions), gestures or the change of body posture and gaze.

Disadvantage(s): Works best for participants who know each other. For participants who do not know each other, difficulties in guessing the other's name may lead to frustration, especially when the reactions of the conversation partner cannot be properly interpreted due to the lack of familiarity.

### A.4 Navigation tasks

Each subject has a partially complete map of the same area, and take turns to explain to the other participants a route through the map. This will include filling in the incomplete sections of the map. Ideally the subjects will need to exchange and confirm a lot of relatively simple information.

The time taken to complete the task may also be used as an indication of the performance.

Good points: Forces all subjects to get involved.

Poor points: Poor for subjective opinion, since subjects tend to rate the complexity of the task.

### A.5 Story with missing parts

In this task the test subjects each get a paper with the same story but with various small sections missing. One participant starts to read the story and the other participants reads the text silently, and interrupts when the reader missed a part of the text. Then the person that noticed that there was a part missing reads that part and continues to read the text until he is interrupted in a similar manner. As the reader continues to read until he hears the other person speak, this conversational task leads to several double talk situations. It is recommended that each conversation lasts between 3 and 5 minutes.

Good points: Extendable to many test subjects.

Poor points: Not suitable for audiovisual tests as the test subjects need to read the stories and are not free to look at their conversation partners.

### A.6 Building blocks task – variation

This is an extension of the small building blocks task described in [ITU-T P.920] for small groups. In the [ITU-T P.920] version one participant has an assembled version of a model and the other a disassembled version. The goal is that both participants have an assembled version at the end. To extend this to a multi-party scenario, all participants have a dissembled version of the model. The instruction manual to build the final model is however in parts distributed to the participant. The goal is again that each participant has an assembled version of the model in the end. Also the

[ITU-T P.920] recommendation of a minimum size of 3 cm × 3 cm × 2 cm should be adjusted to video quality and resolution of the system under test.

Good points: Requires audiovisual interaction.

Poor points: Does not represent a typical conversation.

### A.7 Survival task – variations

This is a task similar to the survival task in [ITU-T P.1301]. Also in this task, the participants are asked to imagine themselves in a dangerous outdoor situation. Contrary to the "rank items"-style of the [ITU-T P.1301] survival task, this task has a quiz style "questions and select answer" interaction. This style makes it easier to use it multiple times with the same participants for a repeated measure design assessment. It shares many properties with the [ITU-T P.1301] survival task: it is a team building exercise in a cooperative consensus seeking task for small groups, thus suitable for participants which are not familiar with each other. It is further also a "knowledge free" task, meaning that it is assumed that participants are not experts in outdoor survival. The task was found to stimulate discussion and was commented as engaging by most participants. In the task a question about a particular outdoor situation is given with three answer options. The correct answer to the question is based on expert opinions and are accompanied with a specific reasoning. Questions and answers can be found under [b-Biech, 2007]. Every participant should answer the questions first alone before the group discussion begins. This is meant to help to assess how well this group of people works as a team: for a working team the participants should achieve at least the amount of points the highest scoring participant had.

Good points: Good for assessing goal oriented ad-hoc group discussions.

Poor points: Participation might be very unequal based on knowledge about the topic.

# Annex B

## Suggestions for delay-critical tasks

(This annex forms an integral part of this Recommendation.)

### B.1 Information exchange tasks

A number of tasks to evaluate the effects of speech delay on communication quality are provided in Appendix I of [ITU-T P.920]. The first 5 of these are simple information exchange tasks and as such very delay sensitive.

1) take turns in counting;
2) take turns reading random numbers aloud as quickly as possible;
3) take turns verifying random numbers aloud as quickly as possible;
4) words with missing letters are completed with letters supplied by the other talker;
5) take turns verifying city names as quickly as possible.

### B.2 Random number verification task timed

This task is a modified version of the number verification task, described in Appendix VIII in [ITU-T P.805]. The essential difference is that test participants are explicitly encouraged to solve the task as quickly as possible, for instance by informing them, that the fastest group of participants in the whole experiment will win a price. This modification has been found to increase the sensitivity [b-Schoenenberg, 2014].

An extension to three parties has been developed for the German language as well [b-Schoenenberg, 2014].

Advantage(s): High sensitivity to delay. Triggers fast interaction between participants.

Disadvantage(s): Lack of naturalness of the task. Participants may try short-cuts by violating the rules of taking turns when reading and confirming the numbers. Thus, test supervisors should listen to the conversations and remind participants not to use short-cuts.

# Bibliography


[b-Berndtsson, 2012] Berndtsson, G., Folkesson, M. and Kulyk, V. (2012), *Subjective quality assessment of video conferences and telemeetings*, Packet video workshop 2012, München, Germany.

[b-Biech, 2007] Biech, E. (2007), *The Pfeiffer book of successful team-building tools: Best of the annuals*. Pfeiffer.

[b-Brady, 1965] Brady, P.T. (1965), *A technique for investigating on-off patterns of speech*, Bell Syst. Tech. J., Vol. 44, No. 1, pp. 1-22.

[b-Brady, 1968] Brady, P.T. (1968), *A statistical analysis of on-off patterns in 16 conversations*, Bell Syst. Tech. J., Vol. 47, No. 1, pp. 73-99.

[b-Brady, 1971] Brady, P.T. (1971), *Effects of transmission delay on conversational behavior on echo-free telephone circuits*, Bell Syst. Tech. J., Vol. 50, No. 1, pp. 115-134.

[b-Clift] Clift, R., *Conversation Analysis*, (Pre-publication draft), pp. 56-83.

[b-Egger, 2012] Egger, S., Schatz, R., Schoenenberg, K., Raake, A. and Kubin, G. (2012), *Same but Different? – Using Speech Signal Features for Comparing Conversational VoIP Quality Studies*, IEEE International Conference on Communications (ICC 2012), Ottawa, Canada.

[b-Gisladottir, 2015] Gisladottir, R., Chwilla, D.J., Levinson, S.C. (2015), *Conversation Electrified: ERP Correlates of Speech Act Recognition in Underspecified Utterances*, PLOS ONE | DOI:10.1371/journal.pone.0120068 March 20.

[b-Guéguin, 2006] Guéguin, M., Le Bouquin-Jeannès, R., Faucon, G. and Barriac, V. (2006), *Towards an objective model of the conversational speech quality*, in Int. Conf. on Acoustics, Speech and Signal Processig (ICASP), IEEE.

[b-Hammer, 2004] Hammer, Florian, Reichl, Peter, and Raake, Alexander (2004), *Elements of interactivity in telephone conversations*. In: Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech, 2004). Jeju Island, Korea, 2004, pp. 1741-1744.

[b-ITU Handbook] *ITU Handbook: Practical procedures for subjective testing* (2011).

[b-Jefferson, 1989] Jefferson, G. (1989), *Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation*. Conversation: An interdisciplinary perspective. Intercommunication series, 3, (pp. 166-196). Roger, Derek (Ed); Bull, Peter (Ed).

[b-Kitawaki, 1991] Kitawaki, N. and K. Itoh, K. (1991), *Pure delay effects on speech quality in telecommunications*, in IEEE Journal on Selected Areas in Communications, Vol. 9, No. 4, pp. 586-593.

[b-Knapp, 2010] Knapp, M.L. and Hall, J.A. (2010), *Nonverbal communication in human interaction*, 7th edition, Wadsworth, Cenage Learning, Boston, USA.

[b-OConnail, 1993] O'Connail, B., Whittaker, S. and Wilbur, J.A. (1993), *Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication*, Human-Computer Interaction, Vol. 8, pp. 389-428.

[b-Raake, 2013] Raake, A., Schoenenberg, K., Skowronek, J. & Egger, S. (2013), *Predicting speech quality based on interactivity and delay*. In: Proc. Interspeech, Lyon, France, August.

[b-Sacks,1974] Sacks, H., Schegloff, E.A., Jefferson, G. (1974), *A simplest systematics for the organisation of turn-taking in conversation*. [Journal] // Language 50. pp. 696-735.

[b-Schmitt, 2014a] Schmitt, M., Gunkel, S., Cesar, P., and Bulterman, D. (2014), *The influence of interactivity patterns on the Quality of Experience in multi-party video-mediated conversations under symmetric delay conditions* in Proceedings of the 3rd International Workshop on Socially-aware Multimedia, New York, NY, USA.

[b-Schmitt, 2014b] Schmitt, M., Gunkel, S., Pablo, C., and Bulterman, D. (2014), *Asymmetric Delay in Video-Mediated Group Discussions* in 6th International Workshop on Quality of Multimedia Experience (QoMEX).

[b-Schoenenberg, 2011] Schoenenberg, K., Raake, A. & Skowronek, J. (2011), *A conversation analytic approach to the prediction of leadership in two to six-party audio conferences.* Third International Workshop on Quality of Multimedia Experience (QoMex), pp. 119-124.

[b-Schoenenberg, 2014] Schoenenberg, K., Raake, A., Egger, S., & Schatz, R. (2014), *On interaction behaviour in telephone conversations under transmission delay*. Speech Communication, 63-64, pp. 1-14.

[b-Schoenenberg, 2014b] Schoenenberg, K., Raake, A., Lebreton, P. (2014), *Conversational Quality and Visual Interaction of Video-Telephony under Synchronous and Asynchronous Transmission Delay*, Sixth International Workshop on Quality of Multimedia Experience (QoMex).

[b-Schoenenberg, 2016] Schoenenberg, K. (2016), *The Quality of Mediated-Conversations under Transmission Delay*. PhD thesis, TU Berlin.

[b-Tam, 2012] Tam, J., Carter, E., Kiesler, S. and Hodgins, J. (2012), *Video increases the perception of naturalness during remote interactions with latency*, in Proc. of CHI'12, New York, NY, USA, pp. 2045-2050.

[b-Tang, 1992] Tang, J.C. (1992), *Why Do Users Like Video? Studies of Multimedia-Supported Collaboration*, Sun Microsystems, Inc., Mountain View, CA, USA.

## SERIES OF ITU-T RECOMMENDATIONS

| | |
|---|---|
| Series A | Organization of the work of ITU-T |
| Series D | General tariff principles |
| Series E | Overall network operation, telephone service, service operation and human factors |
| Series F | Non-telephone telecommunication services |
| Series G | Transmission systems and media, digital systems and networks |
| Series H | Audiovisual and multimedia systems |
| Series I | Integrated services digital network |
| Series J | Cable networks and transmission of television, sound programme and other multimedia signals |
| Series K | Protection against interference |
| Series L | Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant |
| Series M | Telecommunication management, including TMN and network maintenance |
| Series N | Maintenance: international sound programme and television transmission circuits |
| Series O | Specifications of measuring equipment |
| **Series P** | **Terminals and subjective and objective assessment methods** |
| Series Q | Switching and signalling |
| Series R | Telegraph transmission |
| Series S | Telegraph services terminal equipment |
| Series T | Terminals for telematic services |
| Series U | Telegraph switching |
| Series V | Data communication over the telephone network |
| Series X | Data networks, open system communications and security |
| Series Y | Global information infrastructure, Internet protocol aspects and next-generation networks, Internet of Things and smart cities |
| Series Z | Languages and general software aspects for telecommunication systems |

Printed in Switzerland
Geneva, 2016