

Recommendation

## **ITU-T P.836 (03/2023)**

SERIES P: Telephone transmission quality, telephone installations, local line networks

Methods for objective and subjective assessment of speech and video quality

---

**Simulating conversations for the prediction of speech quality**



ITU-T P-SERIES RECOMMENDATIONS

**TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS**

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	P.10–P.19
Voice terminal characteristics	P.30–P.39
Reference systems	P.40–P.49
Objective measuring apparatus	P.50–P.59
Objective electro-acoustical measurements	P.60–P.69
Measurements related to speech loudness	P.70–P.79
Methods for objective and subjective assessment of speech quality	P.80–P.89
Voice terminal characteristics	P.300–P.399
Objective measuring apparatus	P.500–P.599
Measurements related to speech loudness	P.700–P.709
<b>Methods for objective and subjective assessment of speech and video quality</b>	<b>P.800–P.899</b>
Audiovisual quality in multimedia services	P.900–P.999
Transmission performance and QoS aspects of IP end-points	P.1000–P.1099
Communications involving vehicles	P.1100–P.1199
Models and tools for quality assessment of streamed media	P.1200–P.1299
Telemeeting assessment	P.1300–P.1399
Statistical analysis, evaluation and reporting guidelines of quality measurements	P.1400–P.1499
Methods for objective and subjective assessment of quality of services other than speech and video	P.1500–P.1599

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

# Recommendation ITU-T P.836

## Simulating conversations for the prediction of speech quality

### Summary

Recommendation ITU-T P.836 provides a conversation simulation model which is able to simulate realistic conversational behaviour to produce conversations on the semantic level, as well as on the speech signal level. The simulation can replicate conversations with different interactivity patterns and the resulting simulated conversations will reflect changes in this conversation behaviour due to delayed transmission and packet loss. The simulated conversations may be used to predict conversational quality in combination with signal-based or parametric quality prediction models, such as the E-model, e.g., in drive-test scenarios.

### History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.836	2023-03-01	12	<a href="http://handle.itu.int/11.1002/1000/15470">11.1002/1000/15470</a>

### Keywords

Conversation simulation, conversational quality prediction, speech quality.

---

\* 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 received notice of intellectual property, protected by patents/software copyrights, 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 appropriate ITU-T databases available via the ITU-T website at <http://www.itu.int/ITU-T/ipr/>.

© ITU 2023

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

## Table of Contents

	<b>Page</b>
1 Scope .....	1
2 References.....	1
3 Definitions .....	1
3.1 Terms defined elsewhere .....	1
3.2 Terms defined in this Recommendation .....	2
4 Abbreviations and acronyms .....	2
5 Conventions .....	2
6 Conversational behaviour .....	2
6.1 Dialogue acts and concepts.....	3
6.2 Transcriptions .....	3
6.3 Conversation disruptions.....	4
6.4 Turn-taking .....	4
6.5 Conversational interactivity.....	4
7 Simulation of conversational behaviour .....	5
7.1 Incremental simulation framework .....	5
7.2 Turn-taking dialogue manager.....	8
7.3 Turn-taking .....	9
7.4 Conversation disruptions.....	10
8 Application to the conversational quality prediction .....	10
8.1 Quality prediction with the E-model.....	10
Annex A – Python source code.....	11
Appendix I – Example agenda of simulated agents.....	12
Appendix II – Quality predictions from simulated conversations with the E-model .....	14
Bibliography .....	15

Python reference implementation of the conversation simulation.



# Recommendation ITU-T P.836

## Simulating conversations for the prediction of speech quality

### 1 Scope

This Recommendation<sup>1</sup> describes a methodology on how to create a conversation simulation environment that is able to produce conversations that replicate certain characteristics of similar conversations recorded in a conversation test as described in [ITU-T P.800], [ITU-T P.805] and [ITU-T P.804]. The parameters aimed to replicate with the simulation approach are the turn-taking behaviours, the lengths of talk spurts, the interactivity of the conversation, as well as the intents conveyed in each utterance. In contrast to artificial conversational speech described in [ITU-T P.59], the simulation replicates the interactivity patterns of concrete conversation scenarios (e.g., conversation scenarios of [ITU-T P.805] Appendices IV-IX), as well as the impact of degradations such as transmission delay and packet loss on the conversation. The simulated conversations may be used in combination with existing quality models to predict the conversational quality but in itself does not predict the conversational quality.

### 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.107] Recommendation ITU-T G.107 (2015), *The E-model: a computational model for use in transmission planning*.
- [ITU-T G.107.2] Recommendation ITU-T G.107.2 (2023), *Fullband E-model*.
- [ITU-T P.59] Recommendation ITU-T P.59 (1993), *Artificial conversational speech*.
- [ITU-T P.800] Recommendation ITU-T P.800 (1996), *Methods for subjective determination of transmission quality*.
- [ITU-T P.804] Recommendation ITU-T P.804 (2017), *Subjective diagnostic test method for conversational speech quality analysis*.
- [ITU-T P.805] Recommendation ITU-T P.805 (2007), *Subjective evaluation of conversational quality*.

### 3 Definitions

#### 3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

**3.1.1 double talk (DT)** [b-ITU-T P.832]: When near-end and far-end speech occur simultaneously at a given point, typically the terminal under test.

---

<sup>1</sup> This Recommendation includes an electronic attachment with a reference implementation of the conversation simulation in Python (Annex A).

**3.1.2 conversation test** [b-ITU-T P.832]: A subjective test in which two participants have a real-time conversation, as described in Annex A to [ITU-T P.800] and in [b-Telephony].

## **3.2 Terms defined in this Recommendation**

This Recommendation defines the following term:

**3.2.1 conversation disruption:** A turn in a conversation that explicitly or implicitly asks for the retransmission of previous information.

## **4 Abbreviations and acronyms**

This Recommendation uses the following abbreviations and acronyms:

ASR	Automatic Speech Recognition
CD	Conversation Disruption
DT	Double Talk
IU	Incremental Unit
MOS	Mean Opinion Score
MS	Mutual Silence
NLG	Natural Language Generation
NLU	Natural Language Understanding
P-CA	Parametric Conversation Analysis
POTS	Plain Old Telephone System
RNV	Random Number Verification
SA	Speaker A
SAR	Speaker Alternation Rate
SARc	Corrected Speaker Alternation Rate
SB	Speaker B
SCT	Short Conversation Test
TTS	Text-to-Speech
VoIP	Voice over Internet Protocol

## **5 Conventions**

None.

## **6 Conversational behaviour**

For the simulation to be used to predict the speech quality in telephone<sup>2</sup> conversations, a set of conversational behaviour is defined that can be extracted from real conversations and will be modelled in the simulation. This behaviour includes temporal parameters in the form of the timing of turns and measures of the interactivity of the conversation, as well as semantic representation in the form of dialogue acts and concepts that encapsulate the meaning behind the speaking turns. Even

---

<sup>2</sup> "Telephone" here refers to any form of remote speech communication like plain old telephone system (POTS), mobile communication, or over-the-top voice over internet protocol (VoIP) services.

though these parameters can be independently modelled, changes in one behaviour can affect another. For example, an increase in meta-communication caused by misunderstandings due to packet loss can affect the length of utterances and thus, the interactivity of a conversation.

## 6.1 Dialogue acts and concepts

The semantics of the utterances of the dialogue are annotated for both parties of the conversation and represented with dialogue acts and concepts. Dialogue acts represent the general intent conveyed with the utterance, while the concepts reference specific pieces of information that should be requested or provided as part of the conversation. The set of possible dialogue acts are shown in Table 1. They fit the conversation patterns observed in the goal-oriented conversations resulting from conversation tests as described in [ITU-T P.805] and shall be used for every conversation type and scenario. Utterances that may not fit into the dialogue act scheme (e.g., small talk, laughter, or utterances unrelated to the conversation scenario) shall be omitted and will not be simulated.

**Table 1 – List of dialogue acts to encode the semantics of the conversation**

Dialogue act type	Description	Example
greeting	Greeting	"Hello."
goodbye	Farewell	"Goodbye."
provide_info	Providing information	"I want a vegetarian pizza."
provide_partial	Providing parts of information	"My phone number is 0 3 0 ..."
request_info	Requesting information	"What is your address?"
offer_info	Offering information	"Should I give you, my address?"
stalling	Stalling the conversation	"Uhm..." or "Wait a second..."
request_confirm	Request confirmation of information	"You said a vegetarian pizza?"
confirm	Confirming a received information	"Yes." or "Yes, vegetarian."
misunderstanding	Something was not understood	"Could you please repeat that?"
thanks	Giving thanks	"Thank you."
welcome	Receiving thanks	"You're welcome."

The dialogue acts are combined with conversation scenario-specific concepts that represent the information that is being transmitted. For example, for a short conversation test (SCT), the concepts used in combination with the dialogue acts may be *pizza\_type* or *telephone\_number* and for random number verification (RNV) conversations the concepts maybe denoted by *number1*, *number2*, etc. Not every utterance needs to reference a concept. For example, a dialogue act **greeting** may be uttered with the concept *caller\_name* (e.g., "Hello, this is Jeremy Clemens.") or without a concept (e.g., "Hello.").

Each conversation can be represented by a list of dialogue acts and concepts.

## 6.2 Transcriptions

Transcriptions are needed for each dialogue act and concept combination that occurs in the conversation. The transcriptions provide the basis for synthesizing utterances in the simulation and are the key factor in the lengths of utterances in the simulation.

Table 2 shows examples of transcriptions for sets of dialogue acts and concepts.

**Table 2 – Example of transcriptions with dialogue act and concept annotations**

Dialogue act	Concepts	Transcription
greeting	caller_name	"Hello, this is Jeremy Clemens."
greeting		"Good day."
provide_info	pizza_type, pizza_size	"I want a large, vegetarian pizza."
confirm	pizza_type	"Yes, vegetarian."

### 6.3 Conversation disruptions

Conversation disruptions (CD) are defined as a turn in a conversation that explicitly or implicitly asks for retransmission of previous information and hints at a misunderstanding in the directly preceding turn. It is a parameter to measure the effects of bursty packet loss on the structure of the conversation. The packet loss probability and its burstiness are directly correlated with an increase in the conversation disruptions.

However, not every misunderstood word or sentence-fragment results in a conversation disruption (e.g., the interlocutor might reconstruct meaning from the context of the conversation), and a conversation disruption is not necessarily caused by packet loss and could be rooted in other impairments or simply a misunderstanding.

### 6.4 Turn-taking

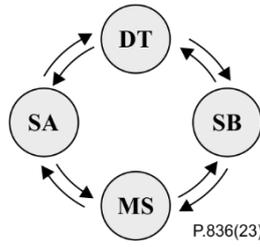
Turn-taking is the set of practices speakers use to organize the conversation and allocate speaking turns. For the simulation of conversations, the turn-taking is analysed and modelled based on the concepts of turn-transitions (i.e., a speaker-change occurs) and turn-continuations (i.e., the turn is continued by the same speaker), which are measured relative to the end of the preceding utterance in seconds [b-Lunsford]. For the turn-transitions, the offsets between the two speech segments are positive if there is a gap between the speaking turns, and negative if the segments overlap (i.e., double talk). Turn-continuations are measured when there is a pause between the utterances of the same speaker and are at least 0.4 seconds long.

The turn-taking behaviour for a conversation is assessed by analysing the turn-transition and turn-continuation timing for each type of conversation or depending on the type of dialogue act that is involved in the transition or continuation of a turn.

### 6.5 Conversational interactivity

The interactivity of a conversation depends on the speed of turn-taking, the length of utterances and the overall setting of the conversation. In order to measure this interactivity, a parametric conversation analysis (P-CA) is used [b-Hammer]. For this analysis, a conversation model based on [ITU-T P.59] is assumed, where the voice activity of both the speakers (referenced as speaker A and B) is used to classify each point in a conversation into four states: speaker A (SA), speaker B (SB), mutual silence (MS), and double talk (DT). The states SA and SB represent the situations in which either speaker A or speaker B is talking only. State MS denotes the case in which nobody is talking, and state DT reflects the cases where both speaker A and speaker B talk at the same time. These four states are used to stochastically model the interaction of a conversation in a Markov process as displayed in Figure 1. Each turn that is taken over by either speaker A or B needs to transition over the states DT or MS.

For a basic analysis of the interactivity of a conversation, the *state probabilities* (i.e., the global probabilities  $P_{MS}$ ,  $P_{DT}$ ,  $P_{SA}$ , and  $P_{SB}$  of each state) and the sojourn times (i.e., the average time in seconds a conversation sojourns in these states) are measured.



**Figure 1 – Four state Markov process modelling the interactions of a conversation**

The main conversational parameter for the assessment of conversational interactivity is the speaker alternation rate (*SAR*), which measures the number of speaker alternations per minute. It is calculated by measuring the number of speaker alternations based on the Markov process and dividing it by the length of the conversation:

$$SAR = \frac{\#SA-MS-SB + \#SB-MS-SA + \#SA-DT-SB + \#SB-DT-SA}{DUR} \quad (6-1)$$

Here, the number of transitions between the state SA and SB (denoted as *SA-MS-SB*, *SB-MS-SA*, *SA-DT-SB*, and *SB-DT-SA*) are counted and divided by the length of the conversation (*DUR*).

The *corrected SAR* (*SAR<sub>C</sub>*) is a delay-based extension of the speaker alternation rate. In contrast to the *SAR*, it considers the added transmission time due to the delay and thus captures the interactivity of a conversation independently of the transmission delay. The definition of the *SAR<sub>C</sub>* is dependent on the side of the conversation on which the interactivity is measured. For speaker A in the conversation, it is defined as:

$$SAR_C^A = \frac{\#SA-MS-SB^A + \#SB-MS-SA^A + \#SA-DT-SB^A + \#SB-DT-SA^A}{DUR - (\#SA-MS-SB^A \cdot 2 \cdot Ta)} \quad (6-2)$$

Here, *SAR<sub>C</sub><sup>A</sup>* is the *SAR<sub>C</sub>* from the perspective of speaker A and *#SA – MS – SB<sup>A</sup>* denotes the number of transitions from the state SA to MS to SB (i.e., a speaker alternation with mutual silence as the transition state) from the perspective of speaker A. The denominator counts all the speaker alternation occurrences from the viewpoint of speaker A. In the divisor, the length of the conversation (*DUR*) in minutes is reduced by the number of speaker changes from person A to B with silence in between, multiplied by two times the one-way transmission delay *Ta* in minutes. Thus, the duration of the conversation is reduced by the full two-way transmission delay overhead of all speaker changes from person A to person B, increasing the calculated speaker alternation rate. The calculation of the *SAR<sub>C</sub><sup>B</sup>* from the perspective of speaker B is analogous to equation 6-2 but subtracts the transitions *#SB – MS – SA<sup>B</sup>* from the duration. The overall *SAR<sub>C</sub>* can be calculated by averaging over both *SAR<sub>C</sub><sup>B</sup>* and *SAR<sub>C</sub><sup>A</sup>*.

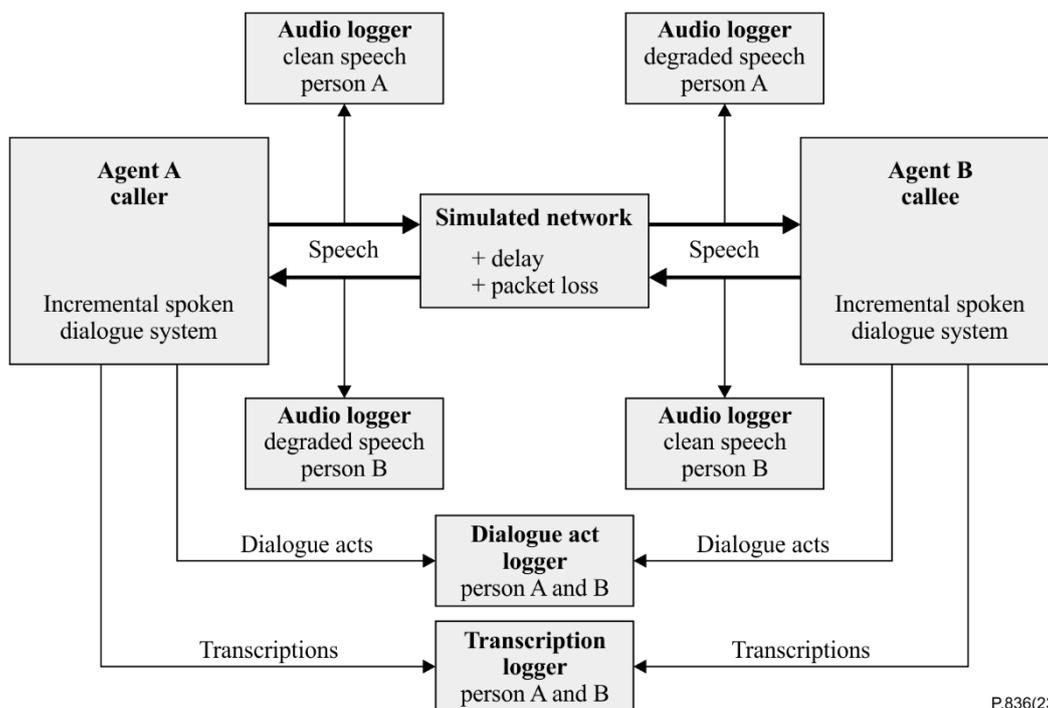
## 7 Simulation of conversational behaviour

In order to accurately reflect the effects of delay and packet loss on a conversation, the conversation simulation needs to process the information on a symbolic level, and it needs to model realistic turn-taking with a focus on timing. The following section defines the core architectural mechanics of the simulation and defines the explicit behaviour for turn-taking, modelling behaviour during packet loss and delay, as well as general dialogue managing.

### 7.1 Incremental simulation framework

The simulation shall be based on two spoken dialogue systems that interact with each other through a simulated voice over internet protocol (VoIP) network. As the simulation requires the two agents to perform timely turn-taking, the dialogue system in the simulation need to interact in an incremental manner [b-Schlangen].

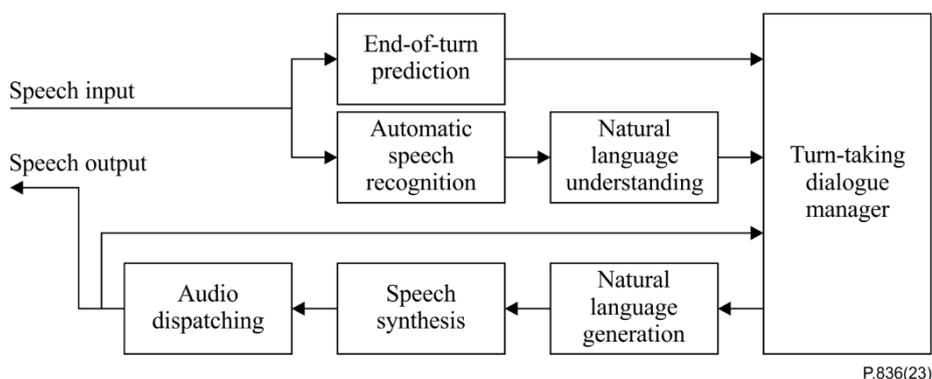
The general simulation approach consists of two spoken dialogue systems, agent A and agent B, that communicate on the speech signal level. Each agent represents one person in the conversation. Agent A takes the role of the caller in each conversation scenario, and Agent B fulfils the role of the callee. The speech signal gets routed in packets (i.e., incremental units) through a simulated telephone network module that is able to delay the arrival of the packets, as well as to replace packets with silence to model zero-insertion packet loss. Logging modules at the non-degraded and degraded end of each agent saves the resulting speech signal to a file for later analysis. Additionally, the dialogue acts and the transcription of the conversation is logged and saved to a file. An overview of the general layout of the simulation network can be seen in Figure 2.



**Figure 2 – Incremental network layout. The two simulated conversation partners are abstracted as spoken dialogue systems, a simulated telephone network introduces impairments, and logging modules save the resulting data to disk**

Because of the incremental nature of this simulation, the agents act unsynchronized and independent of each other. A delay of the speech signal of one agent by the telephone network does not directly affect the mechanics of the other agent. Thus, the agents are only able to adapt their behaviour based on the arrival of the delayed signals that are presented by the telephone network.

Current state-of-the-art models in many areas of spoken dialogue systems, such as speech recognition, end-of-turn prediction, or natural language understanding (NLU), are not on par with human abilities in conversational scenarios. Thus, the simulated telephone network allows for a *side channel* to transmit information about the current state of the interlocutor that might not be easily recoverable from the speech signal alone.



**Figure 3 – Incremental network layout of one agent**

The two spoken dialogue systems of agent A and agent B are constructed as incremental networks shown in Figure 3. The main component of the agents is a *turn-taking dialogue manager*, which orchestrates the taking of turns, dialogue management, and speech dispatching. The turn-taking information is provided by the end-of-turn prediction module that feeds directly from the speech input. The dialogue act and concepts are received by an incremental NLU module that receives the live transcripts from the incremental automatic speech recognition (ASR) module. Once the dialogue manager decides which dialogue acts and concepts should be returned, it provides this information to the natural language generation (NLG) module. There, also a dispatching flag is provided so that the turn-taking dialogue manager can decide when the agent should output the speech and when it should stay silent. The generated text from the NLG module is then synthesized in the text-to-speech (TTS) module. Here, the dispatching flag is still included in the incremental unit (IU) that is being produced. Finally, the audio dispatching module buffers the synthesized speech. When the dialogue manager does not set the dispatching flag, the audio dispatching module produces silence, which is routed to the simulated network. Once the dispatching flag of the dialogue manager is received, it outputs the synthesized speech. The output of the audio dispatching module is also routed to the turn-taking dialogue manager itself. There it is used to monitor the progress of its audio output itself.

The speech recognition and natural language understanding modules of the simulation provide the transcript of the incoming speech and extract the intent and named entities in that transcription. As current speech recognition models do not yield the same accuracy as a human would in the same conversation, the speech recognition and natural language understanding modules used in the simulation do not actually employ a model, but rather rely on the information provided by the side channel of the simulated telephone network. The NLU module receives the transcripts of the speech recognition module. Since the simulated interlocutor has only a limited number of possible utterances that may be outputted, the module looks up the dialogue act and concept associated with the corresponding speech.

Similar to the ASR and NLU modules, the end-of-turn detection module makes use of meta-information provided in the speech IUs provided by the network. For each packet of incoming speech, the module uses a voice activity detection based on a Gaussian mixture model to classify if the interlocutor is speaking or not. If so, the meta-information of the IUs contain the information on how long the turn of the interlocutor will last in seconds. This information is updated for every packet that arrives from the incremental network. The information on whether the interlocutor is speaking and, if so, for how long they will continue to do so is forwarded to the turn-taking dialogue manager.

The NLG module receives a dialogue act and optionally one or multiple concepts from the turn-taking dialogue manager to be turned into natural language text. Additionally, a flag is provided on whether the text that will be synthesized later in the pipeline should be dispatched (i.e., outputted to the interlocutor) or not. This flag is not used in the NLG module, but it is attached to each text IU that is generated by the module.

The speech synthesis (or TTS) module receives the transcripts from the NLG module and produces the corresponding speech. The production of the utterance itself can be done either by synthesizing the speech with a state-of-the-art synthesizer (e.g., Mary TTS) or by using the utterances of already recorded conversations. For this, a database is created, mapping the transcription to the timestamps in the recorded audio files via the dialogue act and turn annotations. Thus, for each utterance that may be produced by the natural language generation (NLG) module, the correct speech file can be loaded, and the position inside the conversation identified and copied to an output speech IU. In order to aid the automatic speech recognition (ASR) module of the other agent, the TTS module adds the text that was synthesized as meta-data to the speech IU. This meta-data will be sent separately through a side channel when the speech is transmitted to the simulated telephone network.

The audio dispatching module receives the speech that should be spoken by the agent, together with a speech dispatching flag that is controlled by the turn-taking dialogue manager. The speech data is stored in a buffer, and depending on whether the dispatching flag of the incoming incremental module is set, the dialogue manager starts to dispatch either the buffered speech or silence at a predefined speed. That way, the audio dispatching module dispatches audio IUs at all times, and alternates between dispatching silence when the dispatching flag is set to *off* by the turn-taking dialogue manager and dispatching the buffered speech when the flag is set to *on*.

For the simulation to be evaluated, data needs to be extracted from every conversation to be analysed. For this, data from the speech, text, and concept layer of the simulation are extracted. An audio recorder module is placed at the outputs of the speech dispatching modules of each agent to capture the clean speech (i.e., speech that has not been altered by the simulated telephone network module) and at the outputs of the telephone network simulator modules to capture the degraded speech. The fullband audio is stored in wave files that can be used for parametric conversation analysis. The transcriptions of the utterances of both agents are stored in a dialogue file. For this, a text recorder module is connected to the output of the NLG module of each agent. The text recorder module only records the uttered turns (turns where the dispatch flag was set) and adds the agent's type (caller or callee) and the timestamp to every transcription. The simulated conversations are also recorded on the dialogue act level. For this, the produced dialogue acts of the turn-taking dialogue manager of each agent are collected by a dialogue act recorder module. Analogous to the text recorder module, the dialogue act recorder adds the agent's type and the timestamp.

## **7.2 Turn-taking dialogue manager**

The turn-taking dialogue manager fulfils the classical task of a dialogue manager by combining the current incoming dialogue act and concepts, the dialogue history, and an agenda of the dialogue to decide what the agent should say in its next turn (in the form of a dialogue act and concepts). However, for the simulation of turn-taking, the agents need to decide when to speak. For this, the module also monitors the progress of the interlocutor's speech to decide when to produce an utterance. For turn-taking, it is also essential for the turn-taking dialogue manager to monitor the progress of its own production of speech. A dialogue manager usually only produces output in the form of dialogue acts and concepts and thus does not have information about when it is producing speech and for how long (as this is usually the task of the speech synthesis module). In this incremental, turn-taking version of a dialogue system, the turn-taking dialogue manager also receives information about its own speech production from the audio dispatching module. This way, it is able to monitor the current speaking status of both sides (i.e., from the interlocutor and itself) to decide when to say what.

To be able to fulfil these tasks, the turn-taking dialogue manager receives three different types of input IUs: the end-of-turn prediction IUs are being received from the end-of-turn prediction module, and dialogue acts and concepts are coming from the NLU module. These two incremental information streams represent the current turn of the interlocutor. The turn-taking dialogue manager also receives the speech IUs of its speech dispatching module to track the progress of its utterances.

The dialogue management part of the module is implemented independently of the turn-taking mechanism, and the dialogue act selection for every dialogue step is realized as an agenda-based dialogue manager and uses the dialogue acts defined in Table 1. In the first step, a stack-based agenda is prepared based on the concepts that should be exchanged during the conversation. These concepts are taken from an agenda that is unique for every concrete conversation scenario that is specific for each conversation scenario and type of agent (i.e., caller and callee). Examples of such agendas can be found in Appendix I. In the agenda, the concepts that need to be requested from and given out to the interlocutor are structured in categories. In each category, the information needs to be transmitted before an agent is able to proceed with the next category (e.g., the transmission of the address to deliver the pizza has to always come after the decision on which pizza to buy). With the concepts from the agenda, a stack is built up that request or offers information based on the order defined in the agenda file. The dialogue acts of greeting and goodbye are added on top and the bottom of the stack. During the conversation, new dialogue acts (e.g., answers to requests for information) are put on top of the dialogue stack. In order to avoid unnecessary dialogue acts, the stack is cleaned after every step by removing dialogue acts that have been made obsolete.

### 7.3 Turn-taking

The turn-transitions and -continuations need to be modelled on the level of the individual speaker alternations and pauses. One model is active during and shortly after the turn of the simulated interlocutor to decide when a turn-transition should occur. The other model is active after the agent's own turn is completed to determine when a turn-continuation should occur. These two competing models then determine when, relative to the (predicted) end of the current turn, a transition or continuation should occur. Depending on which agent starts to speak, the other agent detects the beginning of a new turn, and the turn-taking models of each agent are restarted.

For turn-continuations the following two models are defined:

$$\widehat{C}_A = 0.9251 \cdot (0.8432 + 2.9231 \cdot x^2) + (C_{UI} \cdot 0.2) \quad (7-1)$$

$$\widehat{C}_B = 1.3876 \cdot (0.3607 + 1.2007 \cdot x^2) + (C_{UI} \cdot 0.2) \quad (7-2)$$

where  $x$  is a uniform random variable between 0 and 1 that is selected in an agent for each speaking turn, and  $C_{UI}$  is the number of *unwanted* interruptions that the agent has experienced in the current conversation. Unwanted interruptions are defined as interruptions from the interlocutor that do not occur due to overlaps in turn-transitions (i.e., in the first or last second of an utterance).

Model  $\widehat{C}_B$  shall be used when the current uttered dialogue act of the speaking agent is either `confirm` or `provide_partial`. For all other dialogue acts, the model  $\widehat{C}_A$  shall be used. The output of the model defines the number of seconds an agent shall wait until it continues speaking with a new turn.

For turn-transitions the following two models are defined:

$$\widehat{T}_A = -0.3226 \cdot \log(0.443 \cdot (-1 + \frac{1}{x})) + (C_{CD} \cdot 0.055) \quad (7-3)$$

$$\widehat{T}_B = -0.1598 \cdot \log(0.17 \cdot (-1 + \frac{1}{x})) + (C_{CD} \cdot 0.055) \quad (7-4)$$

Again,  $x$  represents a uniform random variable between 0 and 1 that is selected in an agent for every listening turn. The variable  $C_{CD}$  is the number of conversation disruptions the agent has experienced in the conversation. The model  $\widehat{T}_B$  shall be used for dialogue acts `confirm` or `provide_partial` and the version  $\widehat{T}_A$  shall be used for all other dialogue acts. The resulting output defines the seconds of pause the agent shall wait until taking over the turn from its interlocutor, where negative values represent overlaps (DT), and positive values represent gaps (MS) between speaking turns.

## 7.4 Conversation disruptions

Conversation disruptions affect the overall structure and interactivity of the conversation and are rooted in the misunderstanding of the parts of the utterance due to packet loss. In the simulation, they are modelled by the relationship between the amount of lost speech (due to packet loss) in each utterance and the following occurrence of a conversation disruption. For this, a probability is calculated that the following turn will result in a conversation disruption:

$$\widehat{P}_{CD} = 0.1394 \cdot P_{LS}^2 + 0.1652 \cdot P_{LS} + 0.0035 \quad (7-5)$$

where  $P_{LS}$  is the ratio of lost speech in the previous utterance and  $\widehat{P}_{CD}$  is the probability of a conversation disruption occurring in the following turn. For the agent to access the ratio of lost speech in the previous utterance, the simulated telephone network introduces the status of the packet loss into the side channel. Then, the turn-taking dialogue manager of each agent calculates from the relative number of lost packets of each utterance the percentage of lost speech. With equation 7-5 and a random number generator, it is decided whether the agent *misunderstands* the previous utterance. If this is the case, the dialogue manager inserts a misunderstanding dialogue act with the appropriate concepts on top of the stack.

## 8 Application to the conversational quality prediction

For every simulation, the framework produces full transcripts, semantic summaries (in the form of transmitted dialogue acts), and the clean and degraded speech of the conversation. This data can be used to analyse the resulting simulated conversations and to predict the conversational quality of the conversations.

### 8.1 Quality prediction with the E-model

The E-model uses the *Idd* formula to calculate the impairments related to the echo-free transmission delay [ITU-T G.107.2]. In the narrowband version of the E-model recommended in [ITU-T G.107], two parameters are proposed that account for the influence of delay sensitivity ( $sT$ ) and minimum perceivable delay ( $mT$ ) on the conversation. These parameters are given based on an interactivity class of a conversation, but they can also be calculated based on the interactivity of the conversation.

The two parameters are to be calculated based on [b-Raake]:

$$mT = 436.02 - 71.56 \cdot \log(16.76 + SAR_C) \quad (8-1)$$

$$sT = 0.246 + 0.02 \cdot \exp(0.053 \cdot SAR_C) \quad (8-2)$$

Where the  $SAR_C$  is extracted for every simulated conversation based on the P-CA described in clause 6.5. Then, to predict the  $MOS_{CQO}$  for a specific condition, the E-model predictions of individual conversations are averaged.

In order to obtain stable predictions, at least 30 simulated conversations should be used for the prediction of a single condition.

## **Annex A**

### **Python source code**

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

This Recommendation includes an attachment with a reference implementation of the conversation simulation in Python.

## Appendix I

### Example agenda of simulated agents

(This appendix does not form an integral part of this Recommendation.)

Shown here are the agenda for the simulation of the short conversation test (SCT) scenario 11 and the random number verification (RNV) scenario 1. The information provided in these agenda files corresponds to the information given in the short conversation test scenarios. The information is split into categories denoted by square brackets (e.g., [General]). For each category, information has to be either requested from the interlocutor (when only the name of the information variable is given), or information has to be given to the interlocutor (when the information variable is set to a specific value). If an information is split over multiple lines (e.g., the telephone number), it may (but does not have to) be split over multiple turns.

```
1  [General]
2  callee_name=Pizzeria Roma
3
4  [Reason]
5  reason
6
7  [Additional]
8  num_of_persons
9  pizza_type
10
11 [Offer]
12 pizza_name=Pizza Vegetaria
13
14 [CalleeInformation]
15 toppings=spinach
16   mushrooms
17   tomatoes
18   cheese
19 price=17 Euro
20
21 [CallerInformation]
22 caller_name
23 address
24 telephone
25
26 [Improv]
27 delivery_duration=<improvised>
```

Figure I.1 – Agenda of the callee in the SCT 11 scenario (Pizzeria Roma)

```
1  [General]
2  callee_name
3
4  [Reason]
5  reason=1 large pizza
6
7  [Additional]
8  num_of_persons=2
9  pizza_type=vegetarian
10
11 [Offer]
12 pizza_name
13
14 [CalleeInformation]
15 toppings
16 price
17
18 [CallerInformation]
19 caller_name=Jeremy Clemens
20 address=Gluecksburger Str.
21   41
22   Bochum
23 telephone=0
24   8
25   1
26   1
27   7
28   3
29   4
30   2
31   0
32
33 [Improv]
34 delivery_duration
```

**Figure I.2 – Agenda of the caller in the SCT 11 scenario (Pizzeria Roma)**

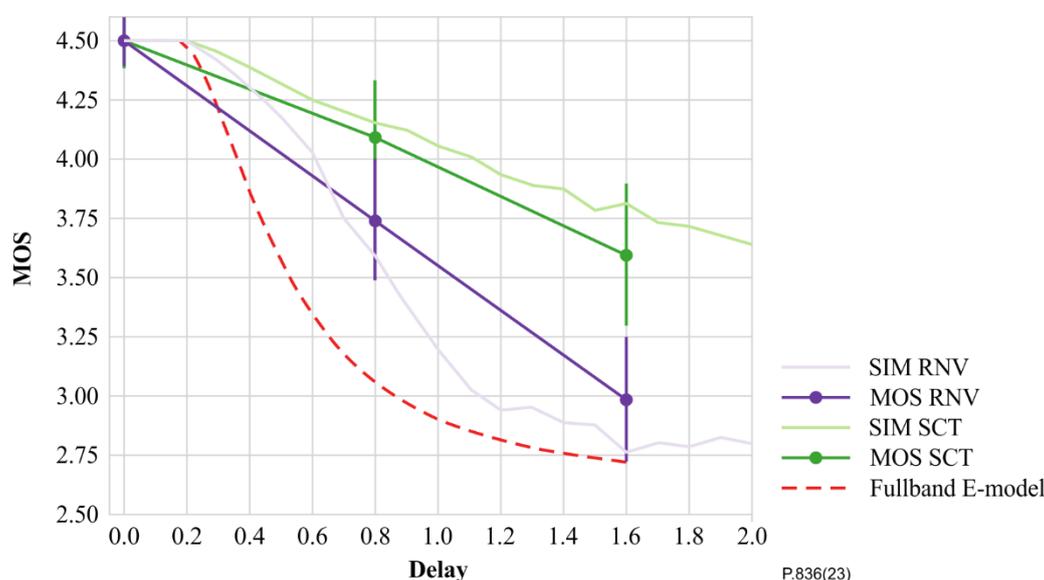
## Appendix II

### Quality predictions from simulated conversations with the E-model

(This appendix does not form an integral part of this Recommendation.)

To evaluate the simulation methodology, 30 SCT and RNV conversations from  $0\text{ ms}$  transmission delay up to  $2\,000\text{ ms}$  transmission delay in  $100\text{ ms}$  steps were simulated, resulting in 1 260 simulated conversations. These were compared against a conversation experiment carried out in accordance with [ITU-T P.805], where 58 participants (age 18-71, 28 of them female) conducted 580 SCT and RNV conversations and delay levels of 0, 800 and 1 600 ms were added.

For every simulated conversation, the  $SARc$  was calculated with equation 6-2 and the  $mT$  and  $sT$  parameters were calculated according to equations 8-1 and 8-2. The fullband E-model was then used to predict the quality of each individual simulated conversation.



**Figure II.1 – Conversational quality mean opinion score (MOS) from the conversation experiment, as well as the predictions from the fullband E-model without simulation (red) and the predictions based on the E-model and simulations (light colours). For the experiment and the simulation, the MOS is split by RNV (purple) and SCT conversations (green)**

Figure II.1 shows the result of the evaluation. Through the difference in interactivity between the simulated SCR and RNV conversations, the prediction is able to replicate the differences in the mean opinion score (MOS) between these two scenarios.

## Bibliography

- [b-ITU-T P.832] Recommendation ITU-T P.832 (2000), *Subjective performance evaluation of hands-free terminals*.
- [b-Hammer] Hammer, F. (2006), *Quality Aspects of Packet-Based Interactive Speech Communication*, PhD Thesis, University of Technology at Graz.  
<<https://theses.eurasip.org/theses/561/quality-aspects-of-packet-based-interactive/>>
- [b-Lunsford] Lunsford, R., Heeman, P.A., and Rennie, E. (2016), *Measuring turn-taking offsets in human-human dialogues*.  
<[https://www.researchgate.net/publication/307889816\\_Measuring\\_Turn-Taking\\_Offsets\\_in\\_Human-Human\\_Dialogues](https://www.researchgate.net/publication/307889816_Measuring_Turn-Taking_Offsets_in_Human-Human_Dialogues)>
- [b-Raake] Raake, A., Schoenenberg, K., Skowronek, J., and Egger-Lampl, S. (2013), *Predicting speech quality based on interactivity and delay*.  
<[https://www.researchgate.net/publication/261262741\\_Predicting\\_Speech\\_Quality\\_Based\\_on\\_Interactivity\\_and\\_Delay](https://www.researchgate.net/publication/261262741_Predicting_Speech_Quality_Based_on_Interactivity_and_Delay)>
- [b-Schlangen] Schlangen, D., and Skantze, G. (2011), *A general, abstract model of incremental dialogue processing*. *Dialogue and Discourse*, 2(1) 83–111.  
<<https://clp.ling.uni-potsdam.de/publications/Schlangen-2011.pdf>>
- [b-Telephonometry] Handbook on Telephonometry (1992), *Measurement methods: telephonometry*.  
<<https://www.itu.int/en/publications/ITU-T/Pages/publications.aspx?parent=T-HDB-MES.2-1993&media=electronic>>





## SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
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
<b>Series P</b>	<b>Telephone transmission quality, telephone installations, local line networks</b>
Series Q	Switching and signalling, and associated measurements and tests
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, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems