

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.807

(02/2016)

SERIES P: TERMINALS AND SUBJECTIVE AND
OBJECTIVE ASSESSMENT METHODS

Methods for objective and subjective assessment of
speech quality

Subjective test methodology for assessing speech intelligibility

Recommendation ITU-T P.807

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
		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 voice services	Series	P.1500

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

Recommendation ITU-T P.807

Subjective test methodology for assessing speech intelligibility

Summary

Recommendation ITU-T P.807 describes a subjective testing methodology for assessing speech intelligibility in communications settings, systems and devices. The method provides a percent correct intelligibility score based on a two-alternative, forced-choice task where the stimulus is one of the two words from a pair of words, i.e., a test item. Half of the test items are rhyming word-pairs (i.e., they differ only in the initial consonant) and half are alliterative word-pairs (i.e., they differ only in the final consonant). The two critical consonants in each test item differ only in a single distinctive feature (see Annex A for a description of distinctive features). In addition to a score for overall intelligibility, the method provides scores for each of six distinctive features: voicing, nasality, sustention, sibilation, graveness and compactness. These scores may be used to diagnose the specific cause of impairments leading to degradation of speech intelligibility.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.807	2016-02-29	12	11.1002/1000/12750

Keywords

Diagnostic assessment of intelligibility, distinctive features, speech intelligibility testing.

* 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/>.

© ITU 2016

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

	Page
1 Scope.....	1
2 References.....	1
3 Definitions	1
3.1 Terms defined elsewhere	1
3.2 Terms defined in this Recommendation.....	1
4 Abbreviations and acronyms	2
5 Conventions	2
6 Description of the ITU-T P.807 testing methodology	2
6.1 Need for an ITU-T intelligibility testing method	2
6.2 Rhyme tests	3
6.3 Intelligibility test method – ITU-T P.807	4
6.4 ITU-T P.807 test results	6
Annex A – ITU-T P.807 and distinctive features	12
Appendix I – Example instructions for the ITU-T P.807 test.....	14
Bibliography.....	17

Recommendation ITU-T P.807

Subjective test methodology for assessing speech intelligibility

1 Scope

This Recommendation describes a subjective testing methodology for assessing speech intelligibility in communications settings, systems and devices. The ITU-T P.807 test methodology has been tested and was found to be appropriate for assessing speech intelligibility of telecommunications systems (e.g., speech codecs), including channel impairments and background noise conditions and for terminals and devices (e.g., handsets, intercom systems). It is designed for use with naive subjects and requires little in the way of specialized equipment or software. The method applies to word intelligibility and does not address intelligibility of phrases or sentences. The method described here applies to North American English. However, the method could be adapted to other languages or dialects taking into account the appropriate set of distinctive features for the language.

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

3 Definitions

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 compactness: Distinctive feature that distinguishes compact phonemes from diffuse phonemes. Compact phonemes are produced by constriction toward the rear of the vocal tract; diffuse phonemes by constriction near the middle. Compact phonemes are characterized by the concentration of spectral energy in the mid-frequency range; diffuse phonemes by the distribution of energy over more-widely separated spectral peaks.

3.2.2 graveness: Distinctive feature that distinguishes grave phonemes from acute phonemes. Grave phonemes are produced by constriction toward the anterior of the vocal tract; acute by constriction in the middle of the tract. Grave phonemes are distinguished among other things by the origin and direction of second-formant transitions. Grave consonants always involve relatively steep upward transitions of the second formant. Acute consonants usually involve downward second-formant transitions, depending on vowel environment and the phoneme involved. In general, grave phonemes are characterized by greater concentration of low-frequency spectral energy than are acute phonemes.

3.2.3 nasality: Distinctive feature that distinguishes nasal phonemes from non-nasal phonemes. Nasals are produced by lowering of the velum, allowing air to escape through the nasal passages; non-nasals by closing the nasal passages. Nasal phonemes are distinguished by relatively pronounced resonances at *circa* 200, 800, and 2200 Hz and by the presence of nulls throughout the frequency range.

3.2.4 sibilation: Distinctive feature that distinguishes sibilated phonemes from non-sibilated phonemes. Sibilants involve extreme constriction of the vocal tract that produces turbulence and high-frequency noise. Sibilant consonants are characterized by higher-frequency noise and greater duration than their non-sibilant counterparts.

3.2.5 sustention: Distinctive feature that distinguishes sustained phonemes from interrupted phonemes. Sustained phonemes are produced by incomplete constriction of the vocal tract; interrupted phonemes by complete constriction of the tract at some point. Sustained phonemes are distinguished by their gradual onset and by the presence of mid-frequency noise, interrupted by their abrupt onset. Sustained phonemes have characteristic durational and high-frequency cues that distinguish them from their interrupted counterparts.

3.2.6 voicing: Distinctive feature that distinguishes voiced phonemes from unvoiced phonemes. Voiced phonemes involve free vibration of the vocal cords; unvoiced phonemes do not. Voiced phonemes are distinguished from their unvoiced counterparts, or cognates, by the presence of periodicity, and, in particular, by the time of onset in periodicity. In voiced consonants, preceding vowels tend to be of greater duration than in the case of unvoiced consonants.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

2AFC	2-Alternative, Forced-Choice
ACR	Absolute Category Rating
AMR-WB	Adaptive Multirate Codec – Wideband
ANSI	American National Standards Institute
DALT	Diagnostic Alliteration Test
DRT	Diagnostic Rhyme Test
HATS	Head And Torso Simulator
IPA	International Phonetic Alphabet
MOS	Mean Opinion Score
MRT	Modified Rhyme Test
RF	Radio Frequency
SNR	Signal to Noise Ratio

5 Conventions

None.

6 Description of the ITU-T P.807 testing methodology

6.1 Need for an ITU-T intelligibility testing method

In recent years, there has been increased interest in testing systems or devices for intelligibility. This is especially relevant for "speech enhancement" techniques and algorithms, e.g., noise

reduction and bandwidth extension, where subjective evaluation has concentrated on speech quality and little is known about the effects of such algorithms on speech intelligibility. The purpose of this Recommendation is to provide such a method.

Most of the subjective testing methodologies, standardized under ITU-T Study Group 12 (SG12), involve the use of relatively large panels of naive listeners, typically a minimum of 32 subjects. This practice has several advantages: 1) there is no need for extensive selection or training of test subjects and 2) the test results can be generalized to the general population of users of the communication systems being tested. The American National Standards Institute (ANSI) standard S3.2 [b-ANSI] specifies that the diagnostic rhyme test (DRT) and modified rhyme test (MRT) use small panels of highly trained and motivated expert listeners to provide stable and reliable results. This has led to the use of panels of eight or fewer test subjects. For practical purposes, this has limited these methods to the relatively few test laboratories that can maintain a panel of trained listeners for routine intelligibility testing.

In addition to the methods that use individual words as stimuli, there are a number of test methods that use longer segments of speech (phrases or sentences) as the test stimuli. These methods have often been deemed impractical for routine testing. The primary criticisms are that sentence-based intelligibility tests are inefficient. The duration of a trial for such longer stimuli, limits the data collection rate to ten or fewer responses per minute of testing, whereas the use of single word-stimuli can raise that response rate by a factor of three. Furthermore, sentence tests typically have little control for the effects of context. With single-word stimuli, there is no context, so context cannot be a confounding factor.

6.2 Rhyme tests

The two most widely used intelligibility tests, the MRT [b-House] and the DRT [b-Voiers], are described in ANSI standard S3.2 [b-ANSI]. Both of these test methods use single syllable stimuli in a multiple-choice task, six choices for the MRT and two choices for the DRT. Both methods express their intelligibility scores in terms of percent correct adjusted for chance (i.e., adjusted for guessing).

The MRT is a test of consonant discrimination in both the initial (rhyming) and the final (alliterative) positions in single-syllable words while the DRT only tests consonants in the initial position. There is, however, a derivative of the DRT that uses the same principles and structure as the DRT, but tests final consonants, i.e., the diagnostic alliteration test (DALT). The approach used here is a combination of DRT items and DALT items to test both initial and final consonants.

Each MRT item includes six response alternatives where the relevant consonants of those alternatives can differ in one to six distinctive features. An analysis of discrimination errors in the results of Miller-Nicely [b-Miller] shows that a vast majority of errors in consonant discrimination occur for single distinctive-feature oppositions. Furthermore, the error rate decreases monotonically with increases in the number of distinctive-feature differences. This finding suggests that the six response alternatives in the MRT are not equally attractive. Alternatives with a high number of distinctive-feature differences from the stimulus are rarely, if ever, confused while the bulk of discrimination errors occur in alternatives with single-feature differences from the stimulus. This means that the MRT is not really a six-choice test and the probability of a discrimination error is dependent on the distribution of distinctive-feature differences among the alternatives. Furthermore, that distribution is different for each MRT item. It follows, then, that the probability of a discrimination error is also different for each of the six words within each item.

For the DRT, however, each item involves a single distinctive-feature difference and discrimination errors can be attributed to that difference alone. An advantage of such un-confounded differences, is the provision of summary scores for each of the six distinctive-features. Those feature scores can be further split into scores for the distinctive-feature in the present state and in the absent state. The set of feature scores, summary, feature-present and feature-absent, provides a DRT profile for the test

condition. These profiles have been used for decades in diagnosing causes of specific impairments hence, the "Diagnostic" feature of the DRT.

6.3 Intelligibility test method – ITU-T P.807

One of the principles of the method described here is the combination of the best attributes of the single-word speech intelligibility tests and the experimental design and naive-subject approach of speech quality testing methods. The test method is designed to provide reliable and valid subjective test results with efficient and cost-effective procedures.

6.3.1 Selection of intelligibility test items

There is much evidence to support the use of a two-choice task for intelligibility-test trials, i.e., a 2-alternative, forced-choice (2AFC) task. The test method uses the previously standardized [b-Voiers] distinctive-feature based approach and selected items from the corpus of DRT items for testing consonants in the initial position and from the corpus of DALT items for testing consonants in the final position. Both the DRT and the DALT include 96 items (i.e., word-pairs), 16 for each of the six distinctive-features. Within those 16 items, there are two items for each of eight vowels with two vowels representing each of the four vowel-quadrants. Table 1 shows the eight vowels represented for each of the six distinctive-features.

Table 1 – Classification by vowel-quadrant of the eight vowels used in the intelligibility test

Vowel Quadrant - place of articulation	IPA Phonetic symbol	Vowel sound	Example DRT item	Example DALT item
High-Back	ɑ	ah	KNOCK-DOCK	HOP-HOT
	ɔ	aw	MOSS-BOSS	LAWS-LOSS
High-Front	o	oh	NOTE-DOTE	GROSS-GROWTH
	u	oo	SHOES-CHOOSE	LOOM-LOON
Low-Back	æ	at	THAN-DAN	RAP-RAT
	ɛ	eh	FENCE-PENCE	EDGE-EGG
Low-Front	ɪ	ee	SHEET-CHEAT	REEF-WREATH
	ɪ	ih	JILT-GUILT	RIM-RIB

The ITU-T P.807 method uses a representative subset of the DRT and DALT items, partitioned into four equivalent groups of items, where each group is presented to a separate panel of naive subjects. Table 2 shows four groups of items, each containing two initial-consonant and two final-consonant items (i.e., word-pairs) for each of the six distinctive-features. For each combination of group and distinctive-feature, the four word-pairs use one vowel from each of the four vowel-quadrants shown in Table 1. For each test condition, one group of items is presented to a separate panel of subjects. Test-item groups are allocated across test conditions in a partially-balanced, randomized-blocks experimental design, the same design recommended for speech quality experiments. See [b-Handbook]. That particular experimental design is optimized to control effects due to time and order of presentation of test-conditions. It is important to control for such "time-order effects" when the effects of learning and fatigue can have significant impact on test scores.

Table 2 – ITU-T P.807 test-items

Distinctive Feature	Sample-1				Sample-2				Sample-3				Sample-4			
	VWL	CONS	Present	Absent	VWL	CONS	Present	Absent	VWL	CONS	Present	Absent	VWL	CONS	Present	Absent
Voicing	ɑ	b/p (i)	BOND	POND	o	v/f (i)	VOLE	FOAL	æ	g/k (i)	GAFF	CALF	ε	d/t (i)	DENSE	TENSE
	u	z/s (i)	ZOO	SUE	ɔ	d/t (i)	DAUNT	TAUNT	l	d3/tj (i)	GIN	CHIN	i	b/p (i)	BEAN	PEEN
	ε	v/f (f)	REV	REF	l	d3/tj (f)	RIDGE	RICH	o	g/k (f)	BROGUE	BROKE	ɑ	d3/tj (f)	HODGE	HOTCH
	i	v/f (f)	SHEAVE	SHEAF	æ	g/k (f)	BAG	BACK	ɔ	z/s (f)	LAWS	LOSS	u	b/p (f)	LUBE	LOOP
Nasality	ε	n/d (i)	NECK	DECK	ɑ	n/d (i)	KNOCK	DOCK	ɔ	m/b (i)	MOSS	BOSS	æ	m/b (i)	MAD	BAD
	i	m/b (i)	MEAT	BEAT	u	m/b (i)	MOOT	BOOT	u	n/d (i)	NOTE	DOTE	l	n/d (i)	NIP	DIP
	ɑ	m/b (f)	BOMB	BOB	ε	m/b (f)	GEM	JEB	æ	n/d (f)	FAN	FAD	ɔ	n/d (f)	BRAWN	BROAD
	u	n/d (f)	NOON	NUDE	i	n/d (f)	SCREEN	SCREED	l	m/b (f)	RIM	RIB	u	n/d (f)	MOAN	MODE
Sustention	æ	ð/d (i)	THAN	DAN	ε	f/p (i)	FENCE	PENCE	ɑ	v/b (i)	VOX	BOX	ɔ	θ/t (i)	THONG	TONG
	l	θ/t (i)	THICK	TICK	i	ʃ/tj (i)	SHEET	CHEAT	u	ʃ/tj (i)	SHOES	CHOOSE	u	ð/d (i)	THOUGH	DOUGH
	ɔ	f/p (f)	GOFF	GAWP	ɑ	v/b (f)	SLAV	SLOB	i	ð/d (f)	SEETHE	SEED	æ	f/p (f)	CALF	CAP
	u	ð/d (f)	LOATHE	LOAD	u	f/p (f)	GOOF	GOOP	ε	ʃ/tj (f)	FLESH	FLETCH	l	v/b (f)	LIVE	LIB
Sibilation	ɔ	d3/g (i)	JAWS	GAUZE	æ	s/θ (i)	SANK	THANK	ε	tj/k (i)	CHAIR	CARE	ɑ	tj/k (i)	CHOP	COP
	o	d3/g (i)	JOE	GO	l	d3/g (i)	JILT	GUILT	i	z/ð (i)	ZEE	THEE	u	tj/k (i)	CHEW	COO
	æ	tj/k (f)	PATCH	PACK	ɔ	s/θ (f)	ROSS	WROTH	ɑ	tj/k (f)	NOTCH	KNOCK	ε	d3/g (f)	EDGE	EGG
	l	s/θ (f)	MISS	MYTH	u	s/θ (f)	GROSS	GROWTH	u	z/ð (f)	SUES	SOOTHE	i	z/ð (f)	BREEZE	BREATHE
Graveness	i	p/t (i)	PEAK	TEAK	l	f/θ (i)	FIN	THIN	u	f/θ (i)	FORE	THOR	u	p/t (i)	POOL	TOOL
	ε	m/n (i)	MET	NET	æ	b/d (i)	BANK	DANK	ɔ	b/d (i)	BONG	DONG	o	w/r (i)	WAD	ROD
	u	m/n (f)	LOOM	LOON	u	b/d (f)	STROBE	STRODE	l	m/n (f)	SHIM	SHIN	i	f/θ (f)	REEF	WREATH
	ɑ	p/t (f)	HOP	HOT	ɔ	f/θ (f)	TROUGH	TROTH	æ	p/t (f)	RAP	RAT	ε	b/d (f)	WEB	WED
Compactness	l	h/f (i)	HIT	FIT	u	ʃ/s (i)	SHOW	SO	u	j/r (i)	YOU	RUE	ε	k/p (i)	KEG	PEG
	æ	g/b (i)	GAT	BAT	ɔ	k/t (i)	CAUGHT	TAUGHT	o	g/d (i)	GOT	DOT	i	j/w (i)	YIELD	WIELD
	u	k/t (f)	OAK	OAT	l	g/b (f)	BIG	BIB	ε	g/d (f)	BEG	BED	ɑ	g/b (f)	NOG	KNOB
	ɔ	g/d (f)	FLOG	FLAWED	æ	ʃ/s (f)	CLASH	CLASS	i	k/p (f)	SEEK	SEEP	u	k/p (f)	DUKE	DUPE

6.3.2 Estimating test parameters and experimental design

The design of any new subjective testing methodology requires the determination of a number of test parameters. For example:

- How many test conditions can be included in an experiment?
- How many talkers should be used for each test-condition?
- How many subjects will be required to provide reliable results?

Similar parameters are recommended for use in designing speech quality tests in the *ITU-T Handbook of subjective testing practical procedures* [b-Handbook].

The Handbook recommends a maximum of two hours per subject for an experiment, including orientation, training, practice and actual test sessions. It also recommends that the ratio of test-time to rest-breaks should be approximately equal. Therefore, the number of test conditions is equal to approximately one hour of testing divided by the amount of time required for testing a single test-condition. The Handbook recommends a minimum of four talkers (two males/two females) per test-condition. The Handbook also recommends a minimum of 32 subjects. With the test-items partitioned into four equivalent groups, the design provides for four panels of eight naive subjects.

A trial duration of 1.75 s was determined to be a comfortable response rate for naive subjects. Therefore, for each combination of test-condition and talker, subjects would hear each of 48 test words (i.e., six distinctive-features × four items per distinctive-feature × two words for each item) for a total of approximately 100 s. per talker and approximately 7 min per test-condition. Therefore, it was determined that an ITU-T P.807 experiment could comfortably accommodate eight test-conditions in a two-hour test session.

The following test parameters were set for an ITU-T P.807 test:

- two-hour test sessions;
- 8 test-conditions;
- 4 listening panels of subjects;
- 4 talkers per test-condition (2 males, 2 females);

- 48 trials per talker per test-condition for each listening panel.

Table 3 shows a sample allocation of test item groups (from Table 2) to talkers for each listening panel and for each of eight test-conditions.

Table 3 – Allocation of test-item groups to talkers by listening panel for each test-condition

Test Cond	Panel-1				Panel-2				Panel-3				Panel-4			
	f1	m1	f2	m2	f1	f2	m1	m2	f1	f2	m1	m2	f1	f2	m1	m2
c01	1	2	3	4	2	3	4	1	3	4	1	2	4	1	2	3
c02	2	3	4	1	3	4	1	2	4	1	2	3	1	2	3	4
c03	3	4	1	2	4	1	2	3	1	2	3	4	2	3	4	1
c04	4	1	2	3	1	2	3	4	2	3	4	1	3	4	1	2
c05	4	3	2	1	3	2	1	4	2	1	4	3	1	4	3	2
c06	3	2	1	4	2	1	4	3	1	4	3	2	4	3	2	1
c07	2	1	4	3	1	4	3	2	4	3	2	1	3	2	1	4
c08	1	4	3	2	4	3	2	1	3	2	1	4	2	1	4	3

Pseudo-randomized presentation orders or playlists are constructed for each listening panel based on the partially-balanced, randomized-blocks experimental design described in [b-Handbook].

6.3.3 Test methods and procedures

6.3.3.1 Listening test environment

Laboratory tests using the ITU-T P.807 methodology should be performed in a listening environment that complies with the minimum requirements specified in [ITU-T P.800].

6.3.3.2 Test sessions

Test duration for each panel of subjects should be a maximum of two hours duration including: instructions, training/practice and multiple sub-sessions for the test itself. Appendix I shows an example set of instructions for one implementation of the ITU-T P.807 methodology.

6.4 ITU-T P.807 test results

ITU-T P.807 test results are presented for each test condition. Results include mean and standard deviation of percent correct scores for the total ITU-T P.807 score¹ and are broken down for the initial and final consonants and for each of the distinctive features. Each percent correct score is "adjusted for guessing" using the formula below:

$$P(c) = [R - W (n-1)] / (R+W)*100$$

where:

R = # correct

W = # incorrect

n = # response alternatives = 2

Figure 1 shows sample ITU-T P.807 results for two standard speech codecs [b-C0256]. Figures 2a-2d show the distinctive feature score profiles for the four test conditions shown in Figure 1.

¹ Total ITU-T P.807 scores are typically based on 128 scores (32 subjects × 4 talkers), each of which are percent correct responses over 48 trials.

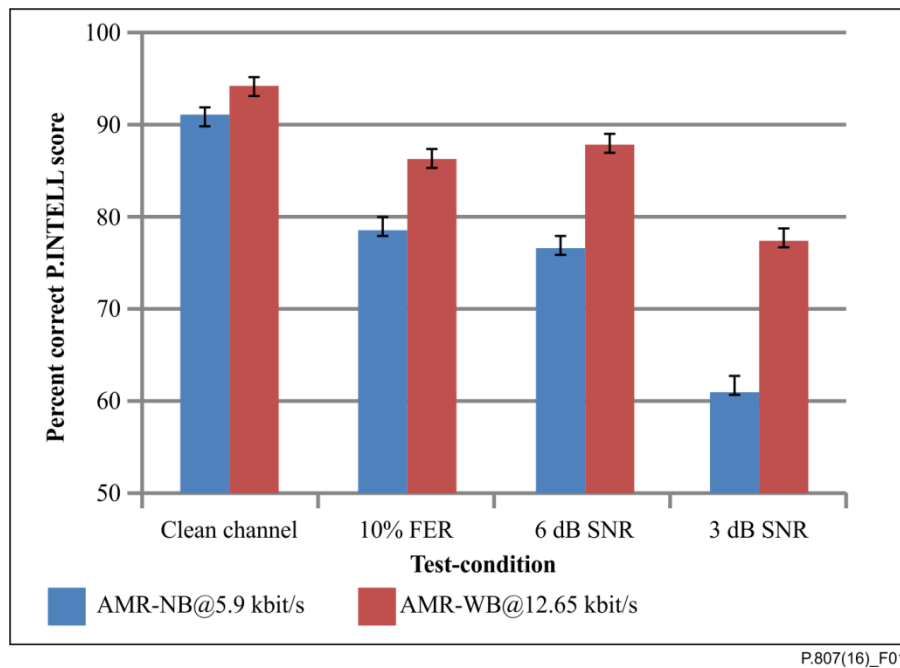


Figure 1 – Sample ITU-T P.807 results for two standard codecs

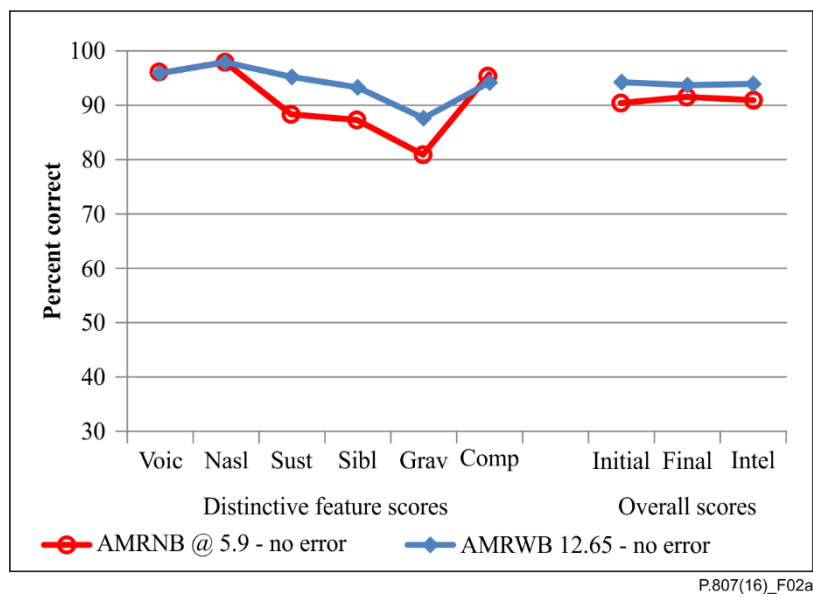


Figure 2a – Score profile for two codecs with no errors

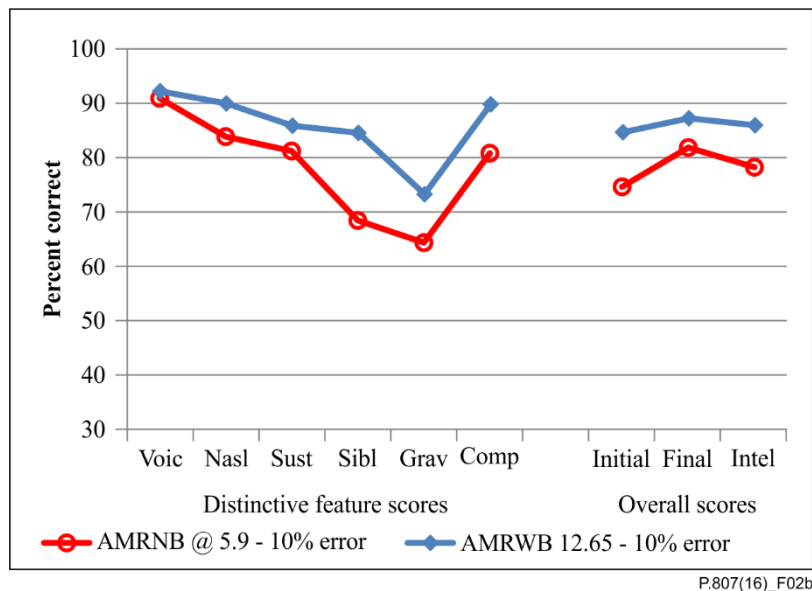


Figure 2b – Score profile for two codecs with 10% errors

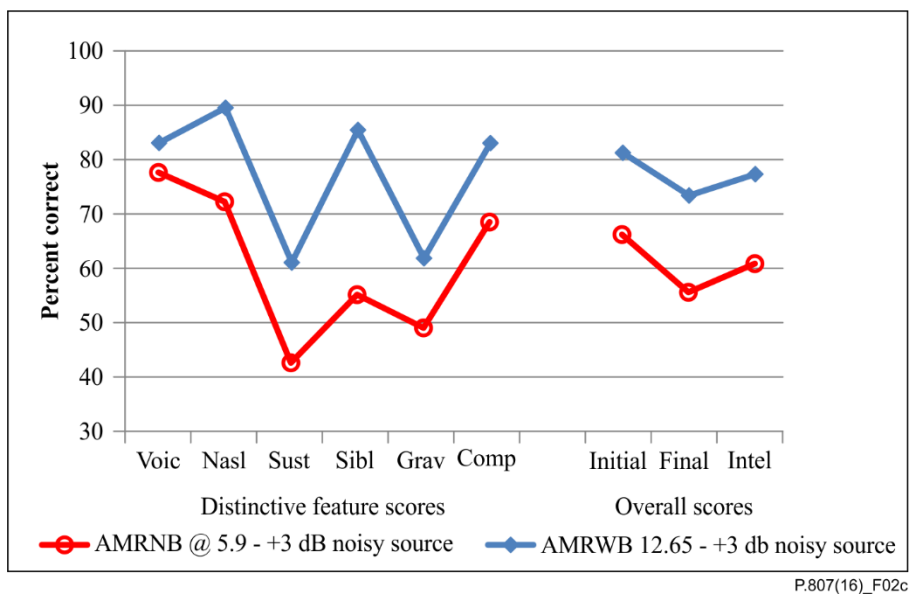


Figure 2c – Score profile for two codecs at 3 dB SNR

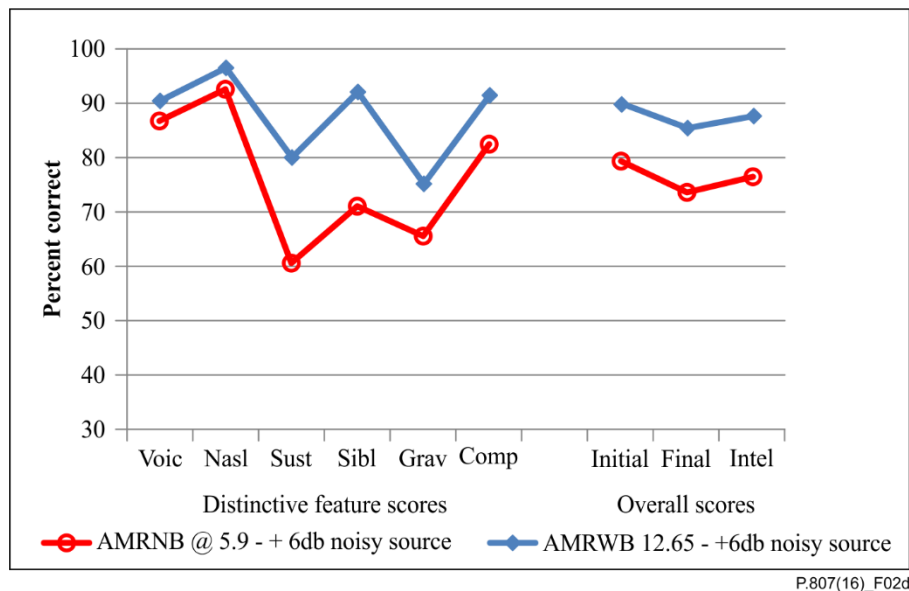


Figure 2d – Score profile for two codecs at 6dB SNR

Figure 3 shows sample ITU-T P.807 results for two commercially available handsets under four background conditions, no noise and three background noises [b-C0296]. Figures 4a-4d show the distinctive feature score profiles for the four test conditions shown in Figure 3.

The receive loudness rating of HS1 (handset 1) was 5dB while the receive loudness rating of HS2 (handset 2) was 5.5 dB. The processing of the test conditions included the following steps. Calls were placed using a network simulator with very good radio frequency (RF) conditions, using adaptive multirate codec – wideband (AMR-WB) with a mode rate of 12.65 kbit/s. The speech signals were injected into the network simulator and the handsets, operating in speakerphone mode, with volume control set to maximum, were placed directly in front of a head and torso simulator (HATS). All recordings were made from both ears of HATS, with free-field equalization. The background noise conditions were generated with reproduction of eight-channel noise recordings using eight loudspeakers, following the method described in [b-ETSI TS 103 224] and using the noise recordings also provided in [b-ETSI TS 103 224].

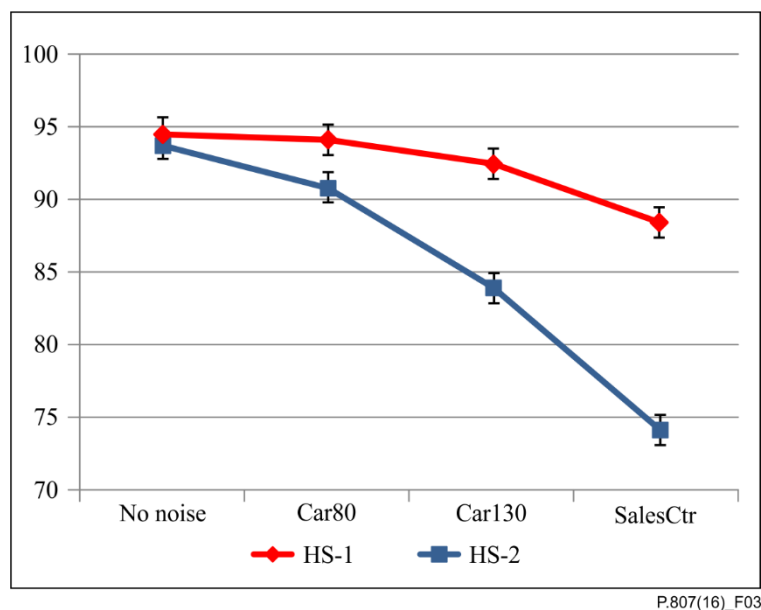


Figure 3 – Sample ITU-T P.807 results for two commercial handsets

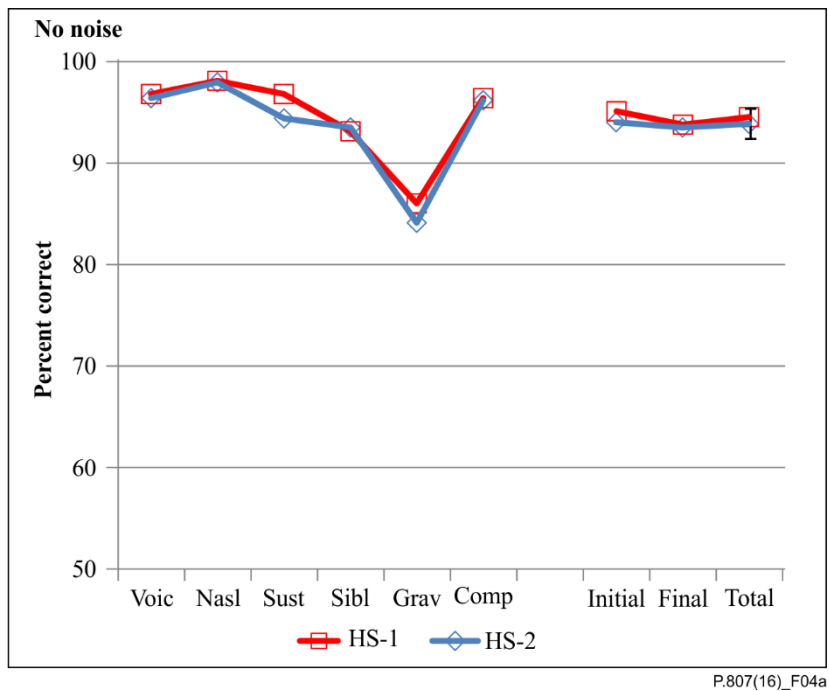


Figure 4a – Profiles for two handsets in clean speech

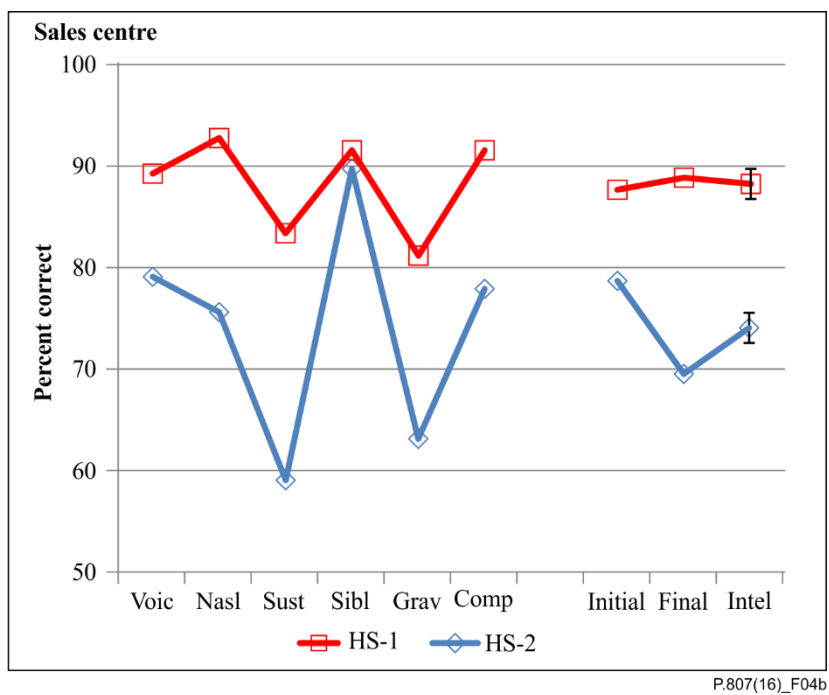
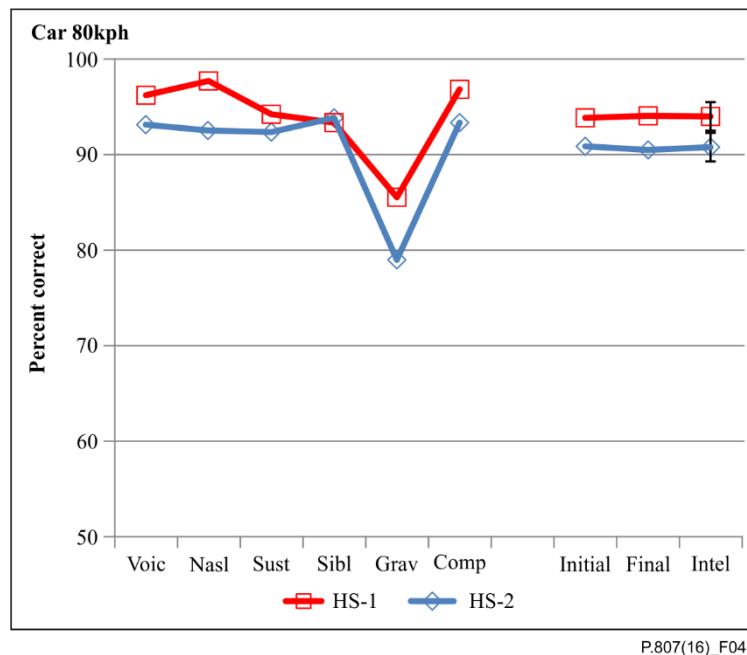
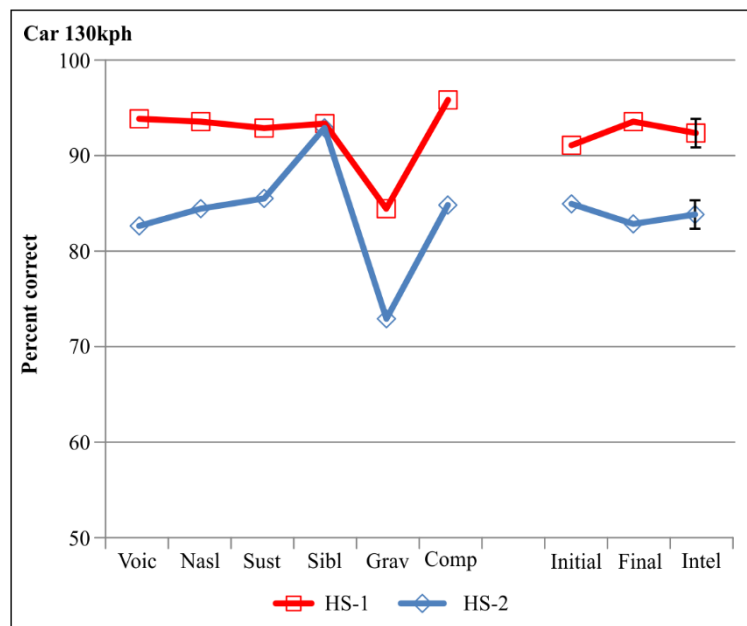


Figure 4b – Profiles for two handsets in sales counter noise



P.807(16)_F04c

Figure 4c – Profiles for two handsets in car 80 kph noise



P.807(16)_F04d

Figure 4d – Profiles for two handsets in car 130 kph noise

A pilot test was conducted [b-C0256] to verify the validity of ITU-T P.807, i.e., "is it measuring what it purports to measure?" ITU-T P.807 test methodology was conducted on the eight test codec conditions illustrated in Figures 1 and 2, using the test procedures described in clause 6.3.3. DRT was also conducted on the same conditions using the procedures described in ANSI standard S3.2 [b-ANSI], including a highly trained panel of expert listeners. The overall intelligibility scores for the two tests showed almost perfect correlation, $r = 0.996$. Furthermore, there was very high agreement for the distinctive feature profile plots for the two methods for all eight test-conditions.

Annex A

ITU-T P.807 and distinctive features

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

The ITU-T P.807 methodology is based on previously standardized intelligibility test methods such as the ANSI standard methodology for measuring intelligibility [b-ANSI] that includes the diagnostic rhyme test or DRT and the related diagnostic alliterative test and the modified rhyme test or MRT. ITU-T P.807 methodology is based on the principle that the intelligibility-relevant information in speech is carried by a small number of distinctive features, such that intelligibility depends most immediately on how well a communication link or device has preserved the acoustical correlates of these features. ITU-T P.807 has adopted this approach from the DRT [b-Voiers], an ANSI standard methodology for measuring intelligibility [b-ANSI]. The DRT was designed, specifically, to measure how well information as to the states of six binary distinctive features: voicing, nasality, sustention, sibilant, graveness and compactness have been preserved by the system or device under test. Table A.1 shows the 23 English consonants and their classification for each of the seven² distinctive features of English consonants. Like the DRT, ITU-T P.807 methodology uses a suite of 96 test-items, where each item is a pair of English words. In half of the test-items, the two words differ only in the initial consonant, i.e., rhyming word-pairs. In the other half, the two words differ only in the final consonant, i.e., alliterative word-pairs. Furthermore, in each test-item the critical consonants, either initial or final, differ only with respect to one of six distinctive features. The listener's task with each item is to judge which of the two words (e.g., **zoo** vs. **sue**, or **bad** vs. **bat**) has been spoken. Incorrect judgments indicate that the system has failed to preserve information contained in the distinctive feature involved. Like most other intelligibility tests in use today, the ITU-T P.807 tests only for the discriminability of consonant phonemes, which carry the bulk of the useful information in speech and are generally more sensitive than vowels to speech degradation. Like the DRT, ITU-T P.807 does not test for the discriminability of vowel-likeness, but does not confound the effects of this feature with those of other features.

The DRT yields a total score, which, under properly controlled conditions, is highly correlated with scores yielded by all other intelligibility tests in use today. The DRT also yields a diversity of diagnostic scores that can be useful in pinpointing specific deficiencies or defects in the system or device under test. With a carefully-selected and monitored panel of eight listeners, the DRT has extremely high resolving power and test-retest reliability. It can resolve differences of less than 1 dB in speech-to-noise ratio.

Distinctive features

The articulatory bases of the six distinctive features are well understood. All voiced phonemes involve free vibration of the vocal cords; unvoiced phonemes do not. Nasals are produced by lowering of the velum, allowing air to escape through the nasal passages; non-nasals by closing the nasal passages. Sustained phonemes are produced by incomplete constriction of the vocal tract; interrupted phonemes by complete constriction of the tract at some point. Sibilants involve extreme constriction of the vocal tract that produces turbulence and high-frequency noise. Grave phonemes are produced by constriction toward the anterior of the vocal tract; acute by constriction in the middle of the tract. Compact phonemes are produced by constriction toward the rear of the vocal tract; diffuse phonemes by constriction near the middle.

2 The DRT and ITU-T P.807 do not test for the discriminability of the vowel-like distinctive feature, but also do not confound the effects of this feature with effects of other features.

Each of the six perceptual distinctive features has multiple acoustical correlates, where the relative saliency of each depends on the phonemic environment and the states of one or more noncritical features. However, some generalizations are possible.

- Voiced phonemes are distinguished from their unvoiced counterparts, or cognates, by the presence of periodicity and, in particular, by the time of onset in periodicity. In voiced consonants, preceding vowels tend to be of greater duration than in the case of unvoiced consonants.
- Nasal phonemes are distinguished by relatively pronounced resonances at *circa* 200, 800, and 2200 Hz and by the presence of nulls throughout the frequency range.
- Sustained phonemes are distinguished by their gradual onset and by the presence of mid-frequency noise; interrupted by their abrupt onset. Sustained phonemes have characteristic durational and high-frequency cues that distinguish them from their interrupted counterparts.
- Sibilant consonants are characterized by higher-frequency noise and greater duration than their non-sibilant counterparts.
- Grave phonemes are distinguished among other things by the origin and direction of second-formant transitions. Grave consonants always involve relatively steep upward transitions of the second formant. Acute consonants usually involve downward second-formant transitions, depending on vowel environment and the phoneme involved. In general, grave phonemes are characterized by greater concentration of low-frequency spectral energy than are acute phonemes.
- Compact phonemes are characterized by the concentration of spectral energy in the mid-frequency range; diffuse phonemes by the distribution of energy over more-widely separated spectral peaks.

Table A.1 – Classification of 23 English consonants by seven distinctive features

	Phoneme*																						
Feature	m	n	v	ð	z	ʒ	dʒ	b	d	g	w	r	l	y	f	θ	s	ʃ	tʃ	p	t	k	h
Voicing	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−
Nasality	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Sustention	−	−	+	+	+	+	−	−	−	−	+	+	+	+	+	+	+	+	−	−	−	−	+
Sibilation	−	−	−	−	+	+	+	−	−	−	−	−	−	−	−	−	+	+	+	−	−	−	−
Graveness	+	−	+	−	−	0	0	+	−	0	+	−	0	0	+	−	−	0	0	+	−	0	0
Compactness	−	−	−	−	−	+	+	−	−	+	−	−	0	+	−	−	−	+	+	−	−	+	+
Vowel-like [#]	−	−	−	−	−	−	−	−	−	−	+	+	+	+	−	−	−	−	−	−	−	−	−

* A plus (+) denotes the nominal or positive state of the feature; a minus (–) denotes the negative state; a zero (0) denotes indifference or neutrality with respect to the feature.

The discriminability of the feature vowel-like is not tested in ITU-T P.807, or in its predecessor, the DRT, but the effects of this feature are not confounded with those of other features.

Appendix I

Example instructions for the ITU-T P.807 test

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

Today you will be involved in an experiment designed to evaluate the intelligibility of speech processed through a number of different telecommunications systems and conditions. The test involves a series of trials where, in each trial, you will be presented a pair of words side-by-side on your computer monitor, and you will hear a single word in your headphones. You will use the computer keyboard to indicate which of the two words displayed on your monitor was spoken by the talker.

- In half of the trials, the two words differ only in their initial consonant. These are "rhyming" word-pairs, for example: **BOB – GOB**, **MOOT – BOOT**, **WIELD – YIELD**.
- In half of the trials, the two words differ only in their final consonant. These are "alliterative" word-pairs, for example: **FAN – FAD**, **LOOM – LOON**, **BEG – BED**.

The trials will be presented in blocks of 24. All of the words within a block will have been spoken by the same talker in the same test condition. Each block begins with a short tone followed by 24 words in two groups: 12 rhyming-word trials and 12 alliterative-word trials. Each word-trial is 1.75 s in duration and each block is 44 s in duration.

You will use three components during the test:

- 1) a set of headphones to listen to the speech materials;
- 2) a computer monitor to display the word-pairs;
- 3) a computer keyboard to register your response for each trial.

Headphones

Your headphones will present the words to both ears. The two earphones are marked with "L" and "R". Put on the headset so that the "L" is on your left ear and "R" is on your right ear. Do not remove your headphones until instructed to do so on your monitor.

Monitor

The computer monitor will show your progress throughout the test, displaying the number of the session, the block and the test word. Figure I.1 shows an example of what your monitor will look like during the test. The first two rows provide information on the subject ID# (222), the session (Practice), the block (1), the total number of blocks in the session (12), the type of word-pair (either Initial Consonant or Final Consonant) and the word # (1) within the block of 24 words. In the middle of the monitor, you are shown a list of four rhyming-word pairs, differing only in the initial consonant. On each trial you will select the word you hear from the two words at the top of the list (i.e., **BOND – POND** in Figure I.1). Your method of selecting a word will be described in the next section. After you have made your response, the word you have selected will be highlighted on your monitor and the list will scroll up i.e., **NECK – DECK** will then be at the top of the list.

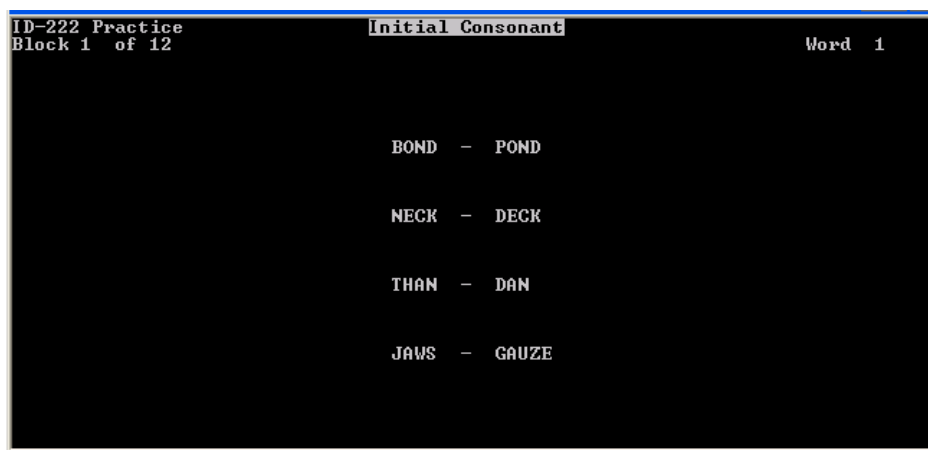


Figure I.1 – Test monitor

Keyboard

In Figure I.2 the keyboard you will use in the test is shown on the left while on the right are the arrow keys you will use during the test to register your choice of word from the word-pair in a trial. During the test itself, the arrow keys are the only active keys on your keyboard.



Figure I.2 – Keyboard and arrow keys

You will use the left-arrow key  to choose the left-hand word from the word-pair and the right-arrow key  to choose the right-hand word from the pair. The word you have chosen will be highlighted on your monitor as illustrated in Figure I.3. Figure I.3 shows the monitor when the left-arrow key was pressed, indicating that the subject chose that the talker said the word "**BOND**". After a short period the set of four word-pairs will scroll up, the next word-pair "NECK – DECK" will be at the top of the list and a new word-pair will be at the bottom.

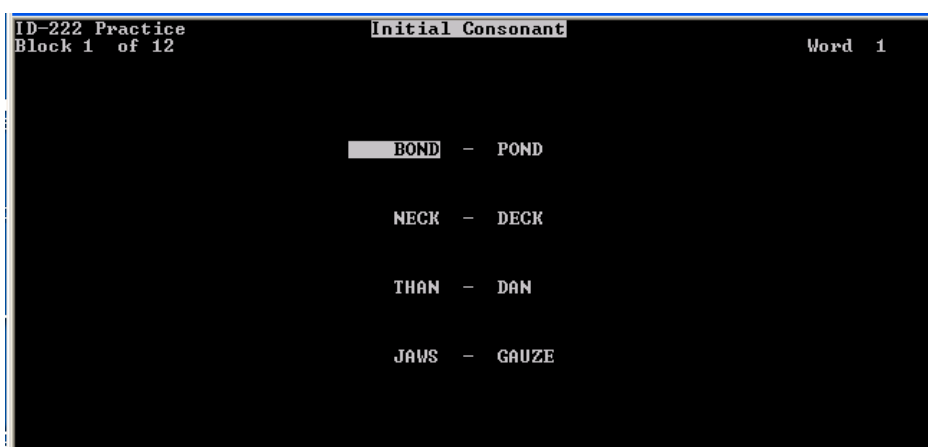


Figure I.3 – The word BOND is highlighted on the monitor

The down-arrow key  is not used in the test, but the up-arrow key  has a special function. If you decide that you have chosen the wrong word on the previous trial, you may press the

up-arrow key and the previous response will be switched to the other word. The previous word-pair will be displayed at the top of the monitor with the other word highlighted. Figure I.4 shows the monitor display if the up-arrow was pressed for the word-pair shown in Figure I.3. Note that the response to the previous word-pair has been switched from "**BOND**" to "**POND**".

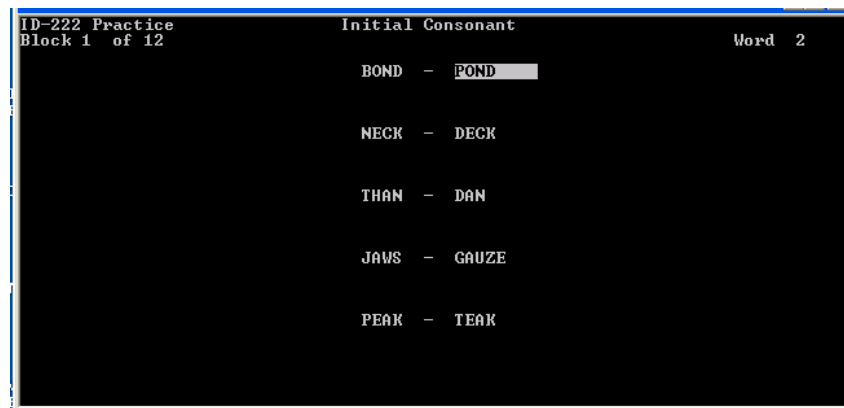


Figure I.4 – The up-arrow is pressed, BOND is corrected to POND

Test sessions

The test administrator will provide you with a subject ID number. It is important that you enter the correct ID number at the beginning of the test or the word-pairs will not be displayed correctly. The test includes four test sessions described below:

- 1) The first session is split into two sections, Practice and Test. The Practice section will include 12 blocks of 24 words and will take about 9 minutes. All of the words within a block will be from the same talker and the same test condition. In the first 4 blocks of Practice, all of the words will be clean, unprocessed speech. For the next 8 blocks, the words will be from each of the 8 test conditions involved in the experiment.
For the second part of Session 1, you will hear the first 8 test blocks, about 6 minutes of testing. At the end of each session, your monitor will instruct you to take off your headphones and leave the listening booth for a short rest break.
- 2) The second session includes 24 blocks and will take about 18 minutes of testing.
- 3) The third session includes 16 blocks and will take about 12 minutes of testing.
- 4) The fourth and final session includes 16 blocks and will take about 12 minutes of testing.

Some of the test conditions will involve clear, unprocessed speech. Others will involve speech in background noise and speech that has been degraded or distorted. The test involves 4 talkers speaking the words for 8 test conditions.

If you have any questions don't hesitate to ask the test administrator now.

Bibliography

- [b-ANSI] ANSI/ASA S3.2 (2009), *Method for Measuring the Intelligibility of Speech over Communication Systems*.
- [b-C0256] ITU-T T13-SG12-C-0256 (2015), *Evaluating Speech Intelligibility – A proposed subjective testing methodology*, Dynastat, Inc., Geneva, Switzerland.
- [b-C0296] ITU-T T13-SG12-C-0296 (2016), *P.INTELL – Method for Evaluating Intelligibility – an Application for Assessing Performance of Wireless Handsets*, Dynastat and Knowles Electronics, Geneva Switzerland.
- [b-CCITT] CCITT/ITU-T Handbook (1992), *Handbook on Telephonometry*.
- [b-ETSI TS 103 224] ETSI TS 103 224 V1.2.1 (2015), *Speech and multimedia Transmission Quality (STQ): A sound field reproduction method for terminal testing including a background noise database*.
- [b-Handbook] ITU-T Handbook (2011), *Handbook of subjective testing practical procedures*.
- [b-House] House, A., Williams, C., Hecker, M.H.L and Kryter, K. (1965), *Articulation testing methods: consonantal differentiation with a closed-response set*, JASA, Vol. 37, No. 1, pp. 158-166.
- [b-Jakobson] Jakobson, R., Fant, G., and Halle, M. (1952), *Preliminaries to speech analysis: the distinctive features and their correlates*, Cambridge, MA: MIT Press.
- [b-Miller] Miller, G.A., and Nicely, P. (1955), *An analysis of perceptual confusions among some English consonants*, JASA, Vol. 27, pp. 338-352.
- [b-Voiers] Voiers, W.D. (1968), *The present state of digital vocoding technique: a diagnostic evaluation*, IEEE Trans. Audio and Electroacoust., Vol. AU-16, No. 2, pp. 275-279.

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