ETSI TR 103 950 V1.1.1 (2023-10)

Speech and multimedia Transmission Quality (STQ); Gender-related aspects of listening quality and effort in speech communication systems

Reference
DTR/STQ-310

Keywords
gender-balanced design, listening effort, listening quality, QoE

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

This Technical Report (TR) has been produced by ETSI Technical Committee Speech and multimedia Transmission Quality (STQ).

The present document describes possible methods of comparing the quality of the transmitted voice for groups of talkers, such as male and female, differing in biological attributes affecting speech (pitch, speaking style, speed, etc.). For the most part, these are biological differences in the human voice-forming apparatus, mostly (but not exclusively) the higher-pitch voices of women. Due to associated socio-cultural implications, such as in the equality of career opportunities between men and women and the non-binary aspects of the distinction between physiological and self-perceived identity, the term "gender" is used further in the present document. Another reason for the choice of this term is the need for the methods described to be applicable also to intentionally or indirectly altered voices (e.g. during hormonal procedures during transition), unrelated to the born physiology of the speaker. See [i.2] and [i.3] for a detailed description and terminology justification.

[i.4] found that fundamental voice pitch is most strongly associated with perceived gender. Other phonetic acoustic or prosodic factors, along with semantics and linguistic style, may play a role. Because in listening tests, the sentence texts are usually constructed to be low-context as best practice and often sorted to be emotionally neutral and randomly assigned to talkers, it seems reasonable to assume that the acoustic phonetics (fundamental pitch, other acoustic factors) are the main variables in the produced speech.

For simplicity, the present document uses:

- (perceived gender) Male = voices rated as masculine-sounding.
- (perceived gender) Female = voices rated as feminine-sounding.

The example spreadsheets are contained in archive tr_103950v010101p0.zip which accompanies the present document.

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Introduction

Perceiving the transmitted speech is a task that puts a certain amount of cognitive load on the human brain. The degree of this load depends on several factors, e.g. the loudness and (coding) distortion of the perceived speech, type and intensity of background noise, quality and accent of the speech, familiarity with the topic of the message, etc. This load also varies between the native and non-native language of the listener. Different levels of such load are manifested in longer duration workloads (e.g. during a work shift) by different levels of overall fatigue, which affects the decrease in the worker's action or decision error rate when performing other concurrent tasks (the so-called parallel-task paradigm).

For technologies used in speech transmission or synthesis, e.g. in telecommunications, radio communications, and machine-to-human communications, the above implies a strong need to optimize human (or synthetic) voice coding to minimize listening effort during communication. Listening Effort (LE) can be assessed by subjective tests following, e.g. Recommendation ITU-T P.800 [i.1], along with Listening Quality (LQ) as specified in Recommendation ITU-T P.800 [i.1]. A natural requirement is that male and female voices are transferred with similar LQ and LE parameters; in other words, the transmission technology, including coding algorithms, frequency filters, or sampling rates, should not privilege one gender over the other to maintain similar working conditions and opportunities for all. Potential misbalance can affect professionals who deploy distant voice communication in their daily duties - e.g. female airport approach control dispatchers or other professionals (female police officers) who are principally disadvantaged by technological aspects of their job - worse voice transmission quality means higher listening effort is needed. It may lead to consequent (subconscious) discomfort of their communication partners. Gender transmission quality misbalance is not surprising for narrow-band or even analogue AM transmissions (still used in aeronautical communications) due to the generally higher pitch region of female voices; however, it is often observed also in contemporary digital wideband or even full-band communications.

1 Scope

The present document addresses the effects of the speaker's gender-related aspects on transmission quality. It provides recommendations on test procedures and implementation means for future technologies dedicated to human speech communication systems, in order to balance transmission quality among genders.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] Recommendation ITU-T P.800: "Methods for subjective determination of transmission Quality".

[i.2] Biemans, M.: "Gender Variation in Voice Quality", LOT Netherlands 2000, ISBN 90-76864-04-7.

[i.3] Pépiot, E.: "Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers", Proc. Speech Prosody 2014, 305-309, doi: 10.21437/SpeechProsody.2014-49.

[i.4] Groll, M.:"Resynthesis of Transmasculine Voices to Assess Gender Perception as a Function of Testosterone Therapy", J Speech Lang Hear Res 2022 Jul 18;65(7):2474-2489. doi: 10.1044/2022-JSLHR-21-00482.

[i.5] Ross, A., Willson, V.L.: "Basic and Advanced Statistical Tests", Springer 2017.

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the following terms apply:

clean reference condition: set of original speech samples, balanced in the number of male and female utterances, usually clean studio recordings

NOTE: Any coding artefacts or background noise neither distorts these speech samples.

CuT conditions: multiple sets of speech samples distorted by CuT

NOTE: They may contain different types and background noise levels, jitter/packet loss artefacts etc. They do not contain any transcoding by CuT AND other non-CuT codec. It may contain samples transcoded by multiple CuT applications.

other reference conditions: usually multiple sets of speech samples, distorted by either reference codec(s) that is (are) not CuT and/or by defined artificial procedures that are not related to CuT (MNRU, ESDRU, and similar)

***p*-value:** probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct

transcoded conditions: set(s) of speech samples, distorted by multiple codings and decodings, using the CuT and other (non-CuT) codec(s)

3.2 Symbols

For the purposes of the present document, the symbols given in Recommendation ITU-T P.800 [i.1] and the following apply:

$DMOS\_F\_CuT_i$	difference between MOS of female voice samples of the *i*-th CuT condition and MOS of female voice samples of the clean reference condition
$DMOS\_M\_CuT_i$	difference between MOS of male voice samples of the *i*-th CuT condition and MOS of male voice samples of the clean reference condition
$STD\_F\_CuT_i$	standard deviation of female voice samples MOS of the *i*-th CuT condition
$STD\_F\_DCuT_i$	standard deviation of $DMOS\_F\_CuT_i$
STD_F_REF	standard deviation of female voice samples MOS of the clean reference condition
$STD\_M\_CuT_i$	standard deviation of male voice samples of the *i*-th CuT condition
$STD\_M\_DCuT_i$	standard deviation of $DMOS\_M\_CuT_i$
STD_M_REF	standard deviation of male voice samples MOS of the clean reference condition
$MOS\_F\_CuT_i$	MOS of female voice samples of the *i*-th CuT condition
MOS_F_REF	MOS of female voice samples of the clean reference condition
$MOS\_M\_CuT_i$	MOS of male voice samples of the *i*-th CuT condition
MOS_M_REF	MOS of male voice samples of the clean reference condition

3.3 Abbreviations

For the purposes of the present document, the abbreviations given in Recommendation ITU-T P.800 [i.1] and the following apply:

CuT	Codec under Test
ESDRU	Energy-based Spatial Distortion Reference Unit
MNRU	Modulated Noise Reference Unit
MOS	Mean Opinion Score
STD	Standard Deviation

4 Evaluation procedure

4.1 Hypothesis statement

For gender-related aspects of listening quality and effort in speech communication systems analysis, the test hypothesis is that samples originating from female and male speakers are transferred with the same quality degradation (*null hypothesis*). The statistical test aims to verify at which confidence level this hypothesis can be rejected.

4.2 Clean reference sample set verification

The original, clean reference sample set quality needs to be verified first for gender neutrality. Therefore, the following parameters are calculated:

- MOS_F_REF and STD_F_REF - arithmetical mean value and the respective standard deviation across all subjective scores of clean reference speech samples spoken by female speakers; and

- MOS_M_REF and STD_M_REF - arithmetical mean value and the respective standard deviation across all subjective scores of clean reference speech samples spoken by male speakers.

The MOS_F_REF and MOS_M_REF usually differ. A suitable statistical test needs to be selected for the understanding the statistical significance of this difference. Assuming common types of reference sample sets (minimum of four speech samples from each of two female and two male speakers, each speech sample being assessed by a minimum of five test subjects), enough subjective scores (typically more than 80 per condition) are acquired. Therefore, a two-tailed independent T-test [i.5] can be used. The clean reference condition quality (mis)balance is described by the T-test conclusion at the selected confidence level (usually alpha = 0,05).

The finding about the quality balance between female and male clean reference sample sets needs to be reported as described in clause 4.4.

NOTE: Subjective results typically vary for reference samples depending on the context of the entire subjective test. Therefore, it is always advisable to validate the reference samples in the context of the test being analysed, even if it has been already analysed in previous tests (e.g. compare reference samples analysis in Example 1 and Example 2 in the Annex A - identical set of reference samples was used in both tests).

4.3 CuT conditions verification

All CuT conditions of the subjective tests need to be considered in the verification. The process is similar to clean reference condition analysis. First, the following parameters are calculated:

- *MOS_F_ CuT$_i$* and *STD_F_ CuT$_i$* - arithmetical mean value and the respective standard deviation across all subjective scores of the *i*-th CuT condition spoken by female speakers are calculated;
- *MOS_M_ CuT$_i$* and *STD_M_ CuT$_i$* - arithmetical mean value and the respective standard deviation across all subjective scores of the *i*-th CuT condition spoken by male speakers are calculated.

Consequently, to compensate for quality differences between female and male clean reference samples, the differences between *MOS_F_ CuT$_i$* and *MOS_F_REF* need to be calculated as:

$$DMOS\_F\_CuT_i = MOS\_F\_CuT_i - MOS\_F\_REF$$

and

$$DMOS\_M\_CuT_i = MOS\_M\_CuT_i - MOS\_M\_REF$$

assuming statistical independence between reference and CuT conditions subjective scores, the standard deviations are calculated as:

$$STD\_F\_DCuT_i = (STD\_F\_CuT_i^2 + STD\_F\_REF^2)^{1/2}$$

and

$$STD\_M\_DCuT_i = (STD\_M\_CuT_i^2 + STD\_M\_REF^2)^{1/2}$$

The difference between *DMOS_F_CuT$_i$* and *DMOS_M_CuT$_i$* shows the difference between the quality degradation of speech samples spoken by female and male speakers. Similarly to the content of clause 4.2, a two-tailed independent T-test [i.5] can be used for the understanding of the statistical significance of this difference.

Each CuT$_i$ condition (mis)balance is described by the T-test conclusion at the selected confidence level (usually alpha = 0,05).

All **other reference conditions** and **transcoded conditions** are excluded from the analysis. Potentially, transcoded conditions inclusion into the test may be done, subject to experimenter experience and particular test circumstances (e.g. if the analysed subjective test is focused on the transcoding performance of the CuT and most of the test conditions are transcoded).

4.4 Test result evaluation and reporting

4.4.1 Overview

An overall evaluation of the subjective test and a conclusion regarding the gender dependence of the tested device is made by testing the null mean hypothesis of the set of differences:

$$DMOS\_F\_CuT_i - DMOS\_M\_CuT_i$$

for *i* covering all **CuT conditions**. All **other reference conditions** and **transcoded conditions** are excluded from the analysis. Potentially, transcoded conditions may be included in the test, subject to experimenter experience and particular test circumstances (e.g. if the analysed subjective test is focused on the transcoding performance of the CuT and most of the test conditions are transcoded).

Unlike the tests of statistical significance in clauses 4.2 and 4.3 (where the normality of the random majority being tested is guaranteed by a central limit theorem) for the set of values DMOS_F_CuTi - DMOS_M_CuTi it is appropriate to investigate its level of agreement with a normally distributed random variable using a suitable statistical test, e.g. Shapiro-Wilk [i.5]. If its distribution is close to the normal distribution, two-tailed paired T-test [i.5] can be used for pairs of DMOS_F_CuTi and DMOS_M_CuTi. Otherwise, a suitable non-parametric test of statistical significance, e.g. Wilcoxon signed rank test [i.5] should be used.

4.4.2 Reported parameters

A final report should contain the following parameters:

a) Numbers of female and male speakers used to create the speech samples. Overall number of speech samples spoken by female speakers per condition. Overall number of speech samples spoken by male speakers per condition.

b) The number of subjective votes collected for each speech sample during the analysed test.

c) Clean reference sample set verification results as per clause 4.2, containing *MOS_F_REF*, *STD_F_REF*, *MOS_M_REF*, *STD_M_REF*, and the T-test *p*-value.

d) CuT conditions verification results as per clause 4.3, containing $DMOS\_F\_CuT_i$ and $DMOS\_M\_CuT_i$ values and respective T-test *p*-values.

e) Overall evaluation result as per clause 4.4.1, containing the information about the statistical test used, significance level selected (usually 0,05), and the resulting *p*-value.

Annex A:
Evaluation examples

A.1 Evaluation example 1

Reporting of the experiment result analysis example (full raw data available in the spreadsheet Example_1 included in the archive tr_103950v010101p0.zip which accompanies the present document.).

a) Number of female speakers: two
Number of male speakers: two
Overall number of speech samples spoken by female speakers per condition: 12
Overall number of speech samples spoken by male speakers per condition: 12

b) Number of subjective votes collected for each speech sample during the analysed test: four

c) Clean reference sample set verification results:

- *MOS_F_REF* = 4,458, *STD_F_REF* = 0,713
- *MOS_M_REF* = 4,604, *STD_M_REF* = 0,644
- T-test *p*-value: p = 0,298.

d) CuT conditions verification results (p = 0,05)

Table A.1-1: Conditional MOS degradation of samples originating from male and female samples and their comparison by T-test (shortened)

	DMOS_M_CuT$_i$	***DMOS_F_CuT$_i$***	***T-test p-value***
c02	0,083	0,292	0,165
c03	0,104	0,458	*0,017*
c04	0,125	0,250	0,366
c05	0,396	0,688	*0,046*
c06	0,229	0,396	0,274
c07	1,167	1,250	0,643
c08	0,313	0,667	*0,014*
c09	0,438	0,667	0,131
c10	0,229	0,729	*0,001*
c11	0,521	0,958	*0,004*
…	…	…	…
c53	0,771	1,250	*0,003*
c54	1,167	1,896	*0,000*
c55	0,833	1,271	*0,008*
c56	1,896	2,083	0,287
c57	0,896	0,813	0,653
c58	1,625	1,729	0,569
c59	2,417	2,354	0,689

e) Overall evaluation result:

The Shapiro-Wilk test [i.5] did not show a significant departure (of set of conditions c02…c59) from normality, *W(58)* = 0,98, p = 0,541 (calculated using https://www.statskingdom.com/shapiro-wilk-test-calculator.html).

Table A.1-2: Final MOS degradation of all samples originating from male and female speakers and their comparison by T-test

MEAN	**-0,249**
STD	**0,161**
No of conditions	**58**
T-test *p*-value	**1,6131E-05**

Conclusion: The CuT is gender-aspect MISBALANCED, p = 0,000016.

A.2 Evaluation example 2

Reporting of the experiment result analysis example (full raw data available in the spreadsheet Example_2 included in the archive tr_103950v010101p0.zip which accompanies the present document.).).

a) Number of female speakers: two
Number of male speakers: two
Overall number of speech samples spoken by female speakers per condition: 12
Overall number of speech samples spoken by male speakers per condition: 12

b) Number of subjective votes collected for each speech sample during the analysed test: four

c) Clean reference sample set verification results:

- *MOS_F_REF* = 4,792, *STD_F_REF* = 0,582
- *MOS_M_REF* = 4,896, *STD_M_REF* = 0,371
- T-test *p*-value: p = 0,301.

d) CuT conditions verification results (p = 0,05)

Table A.2-1: Conditional MOS degradation of samples originating from male and female speakers and their comparison by T-test (shortened)

	DMOS_M_CuT$_i$	***DMOS_F_CuT$_i$***	***T-test p-value***
c03	0,021	0,021	1,000
c04	-0,021	-0,063	0,651
c05	0,313	0,208	0,397
c06	0,500	0,146	*0,004*
c07	0,188	0,167	0,849
c08	0,083	0,000	0,388
c09	0,125	0,229	0,355
c10	0,104	-0,063	0,081
c11	0,521	0,667	0,301
c12	0,083	0,063	0,836
…	…	…	…
c34	0,083	-0,063	0,133
c35	0,104	0,396	*0,013*
c36	0,563	0,688	0,336
c37	0,521	0,750	0,103
c38	1,250	1,167	0,558
c39	0,042	0,000	0,682
c40	0,229	0,417	0,131

e) Overall evaluation result:

The Shapiro-Wilk test [i.5] did not show a significant departure (of the set for conditions c02…c40) from normality, *W(38)* = 0,96, p = 0,195 (calculated using https://www.statskingdom.com/shapiro-wilk-test-calculator.html).

Table A.2-2: Final MOS degradation of all samples originating from male and female speakers and their comparison by T-test

MEAN	**0,014**
STD	**0,122**
No of conditions	**38**
T-test *p*-value	**0,810**

Conclusion: The CuT is gender-aspect BALANCED, $p = 0,810$.

History

Document history		
V1.1.1	October 2023	Publication