الاتحاد الدولي للاتصالات

ITU-R

قطاع الاتصالات الراديوية في الاتحاد الدولي للاتصالات

التوصية ITU-R BS.1534-2

(2014/06)

طريقة التقييم الشخصي لسويات الجودة المتوسطة للأنظمة السمعية

السلسلة BS

الخدمة الإذاعية (الصوتية)

الاتحاد الدولي للاتصالات ITU

150 ITU 1865-2015

تمهيد

يضطلع قطاع الاتصالات الراديوية بدور يتمثل في تأمين الترشيد والإنصاف والفعالية والاقتصاد في استعمال طيف الترددات الراديوية في جميع خدمات الاتصالات الراديوية، بما فيها الخدمات الساتلية، وإجراء دراسات دون تحديد لمدى الترددات، تكون أساساً لإعداد التوصيات واعتمادها.

ويؤدي قطاع الاتصالات الراديوية وظائفه التنظيمية والسياساتية من خلال المؤتمرات العالمية والإقليمية للاتصالات الراديوية وجمعيات الاتصالات الراديوية بمساعدة لجان الدراسات.

سياسة قطاع الاتصالات الراديوية بشأن حقوق الملكية الفكرية (IPR)

يرد وصف للسياسة التي يتبعها قطاع الاتصالات الراديوية فيما يتعلق بحقوق الملكية الفكرية في سياسة البراءات المشتركة بين قطاع تقييس الاتصالات وقطاع الاتصالات الراديوية والمنظمة الدولية للتوحيد القياسي واللجنة الكهرتقنية الدولية (ITU-T/ITU-R/ISO/IEC) والمشار إليها في الملحق 1 بالقرار ITU-R 1. وترد الاستمارات التي ينبغي لحاملي البراءات استعمالها لتقديم بيان عن البراءات أو للتصريح عن منح رخص في الموقع الإلكتروني http://www.itu.int/ITU-R/go/patents/en حيث يمكن أيضاً الاطلاع على المبادئ التوجيهية الخاصة بتطبيق سياسة البراءات المشتركة وعلى قاعدة بيانات قطاع الاتصالات الراديوية التي تتضمن معلومات عن البراءات.

سلاسل توصيات قطاع الاتصالات الراديوية

(يمكن الاطلاع عليها أيضاً في الموقع الإلكتروني http://www.itu.int/publ/R-REC/en)

السلسلة	العنوان
BO	البث الساتلي
BR	التسجيل من أجل الإنتاج والأرشفة والعرض؛ الأفلام التلفزيونية
BS	**الخدمة الإذاعية (الصوتية)**
BT	الخدمة الإذاعية (التلفزيونية)
F	الخدمة الثابتة
M	الخدمة المتنقلة وخدمة الاستدلال الراديوي وخدمة الهواة والخدمات الساتلية ذات الصلة
P	انتشار الموجات الراديوية
RA	علم الفلك الراديوي
RS	أنظمة الاستشعار عن بعد
S	الخدمة الثابتة الساتلية
SA	التطبيقات الفضائية والأرصاد الجوية
SF	تقاسم الترددات والتنسيق بين أنظمة الخدمة الثابتة الساتلية والخدمة الثابتة
SM	إدارة الطيف
SNG	التجميع الساتلي للأخبار
TF	إرسالات الترددات المعيارية وإشارات التوقيت
V	المفردات والمواضيع ذات الصلة

ملاحظة: *تمت الموافقة على النسخة الإنكليزية لهذه التوصية الصادرة عن قطاع الاتصالات الراديوية بموجب الإجراء الموضح في القرار ITU-R 1.*

النشر الإلكتروني
جنيف، 2015

التوصية ITU-R BS.1354-2

طريقة التقييم الشخصي لسويات الجودة المتوسطة للأنظمة السمعية

(المسألة ITU-R 62/6)

(2014-2003-2001)

مجال التطبيق

تصف هذه التوصية طريقة للتقييم الشخصي للجودة السمعية المتوسطة. وتعكس هذه الطريقة العديد من جوانب التوصية ITU-R BS.1116 وتستعمل نفس مقياس تحديد الدرجات المستعمل في تقييم جودة الصورة (أي التوصية ITU-R BT.500).

وتم بنجاح اختبار الطريقة المسماة "اختبار متعدد الحوافز بمرجع ومرتكز مخجوبين (MUSHRA)". وأظهرت هذه الاختبارات أن طريقة MUSHRA مناسبة لتقييم الجودة السمعية المتوسطة وتحقق نتائج دقيقة وموثوقة.

الكلمات الرئيسية

اختبار الاستماع، والأصوات المصطنعة، والجودة السمعية المتوسطة، والتشفير السمعي، والتقييم الشخصي، والجودة السمعية.

إن جمعية الاتصالات الراديوية للاتحاد الدولي للاتصالات،

إذ تضع في اعتبارها

أ) أن التوصيات ITU-R BS.1116، وITU-R BS.1284، وITU-R BT.500، وITU-R BT.710، وITU-R BT.811 وكذلك التوصيات ITU-T P.800، وITU-T P.810، وITU-T P.830 قد حددت طرائق للتقييم الشخصي لجودة الأنظمة السمعية والفيديوية والكلامية؛

ب) أن أنواعاً جديدة من خدمات تقديم المحتوى مثل تدفق المحتويات السمعية على الإنترنت أو الأجهزة بأشباه الموصلات، أو الخدمات الساتلية الرقمية، أو أنظمة الموجات القصيرة والمتوسطة الرقمية، أو تطبيقات الوسائط المتعددة المتنقلة يمكن أن تعمل بجودة سمعية متوسطة؛

ج) أن المقصود من التوصية ITU-R BS.1116 هو تقييم الانحطاط الضعيف وأنها لا تناسب تقييم أنظمة ذات جودة سمعية متوسطة؛

د) أن التوصية ITU-R BS.1284 لا تعطي أي درجات مطلقة لتقييم الجودة السمعية المتوسطة؛

هـ) أن إدماج المرتكزات المناسبة وذات الصلة في عملية الاختبار يسمح باستعمال مستقر لمقياس التقييم الشخصي؛

و) أن التوصيات ITU-T P.800، وITU-T P.810، وITU-T P.830 تركز على إشارات الكلام في بيئة هاتفية وثبت عدم كفايتها لتقييم الإشارات السمعية في بيئة إذاعية؛

ز) أن استعمال طرائق الاختبار الشخصية المقيّسة مهم لتبادل بيانات الاختبار ومواءمتها وتقييمها بالشكل الصحيح؛

ح) أن خدمات الوسائط المتعددة الجديدة قد تتطلب تقييماً مشتركاً للجودة السمعية والفيديوية؛

ط) أن اسم MUSHRA (اختبار متعدد الحوافز بمرجع ومرتكز مخجوبين) كثيراً ما يستعمل بطريقة خاطئة للإشارة إلى الاختبارات التي لا تستعمل مرجع ومرتكزات؛

ي) أن المرتكزات يمكن أن تؤثر في نتائج الاختبار ومن المرغوب فيه أن تكون المرتكزات مشابهة للأصوات المصطنعة التي يجري اختبارها،

توصي

1 باستعمال إجراءات الاختبار والتقييم الواردة في الملحق 1 من هذه التوصية في التقييم الشخصي للجودة السمعية المتوسطة،

توصي كذلك

1 بمواصلة الدراسات المتعلقة بالمرتكزات التي تتميز بخصائص مظاهر الانخطاط التي تُصادف في أحدث الأنظمة السمعية وتحديث هذه التوصية لتشمل المرتكزات الجديدة عندما تكون ملائمة.

الملحق 1

1 مقدمة

تصف هذه التوصية طريقة للتقييم الشخصي للجودة السمعية المتوسطة. وتعكس هذه الطريقة العديد من جوانب التوصية ITU-R BS.1116 وتستعمل نفس مقياس تحديد الدرجات المستعمل في تقييم جودة الصورة (أي التوصية ITU-R BT.500).

وتم بنجاح اختبار الطريقة المسماة "MUSHRA" اختصاراً لعبارة اختبار متعدد الحوافز بمرجع ومرتكز محجوبين. وأظهرت هذه الاختبارات أن طريقة MUSHRA مناسبة لتقييم الجودة السمعية المتوسطة وتحقق نتائج دقيقة وموثوقة [2؛ و4؛ و3].

وتشمل هذه التوصية الأقسام والمرفقات التالية:

القسم 1: مقدمة

القسم 2: مجال التطبيق، ودوافع الاختبار، والغرض من الطريقة الجديدة

القسم 3: التصميم التجريبي

القسم 4: اختيار المقيّمين

القسم 5: طريقة الاختبار

القسم 6: النعوت

القسم 7: مادة الاختبار

القسم 8: ظروف الاستماع

القسم 9: التحليل الإحصائي

القسم 10: تقرير الاختبار وعرض النتائج

المرفق 1 (معياري): التعليمات التي يتعين إعطاؤها للمُقيّمين

المرفق 2 (إعلامي): ملاحظات توجيهية بشأن تصميم السطح البيني للمستعمل

المرفق 3 (معياري): وصف للمقارنة الإحصائية غير المعْلَمية بين عينتين باستعمال تقنيات إعادة اختيار العينة وطرائق محاكاة مونت كارلو

المرفق 4 (إعلامي): ملاحظات توجيهية بشأن التحليل الإحصائي المعْلَمي

المرفق 5 (إعلامي): متطلبات السلوك الأمثل للمرتكزات

2 مجال التطبيق ودوافع الاختبار والغرض من الطريقة الجديدة

من المسلّم به أن اختبارات الاستماع الشخصية لا تزال أكثر الطرائق موثوقيةً لقياس جودة الأنظمة السمعية. وتوجد طرائق جرى وصفها وإثباتها بشكل جيد لتقييم الجودة السمعية في قمة مدى الجودة وعند أدنى درجاته.

وتستعمل التوصية ITU-R BS.1116 – طرائق التقييم الشخصي للانحطاط الضعيف في الأنظمة السمعية، بما في ذلك الأنظمة الصوتية متعددة القنوات، لتقييم الأنظمة السمعية ذات الجودة العالية التي توجد بها مظاهر الانحطاط الضعيف. غير أنه توجد تطبيقات تكون فيها الجودة السمعية الأقل مقبولة أو لا سبيل إلى تفاديها. وأدت التطورات السريعة في استعمال الإنترنت لتوزيع وإذاعة المادة السمعية، حيث معدل البيانات محدود، إلى نوع من التنازلات إزاء الجودة السمعية. والتطبيقات الأخرى التي يمكن أن تحتوي على جودة سمعية متوسطة هي الإشارة الرقمية المشكلة بالاتساع (digital AM) (مثل الراديو الرقمي العالمي (DRM)، والإذاعة الساتلية الرقمية، ودارات التعليق في الراديو والتلفزيون، والخدمات السمعية التي تقدم عند الطلب، والمحتويات السمعية على خطوط المراقمة). وطريقة الاختبار المعرّفة في التوصية ITU-R BS.1116 غير مناسبة تماماً لتقييم هذه الأنظمة السمعية ذات الجودة المتدنية [4] بسبب ضعف قدرتها على التمييز بين الاختلافات الصغيرة في الجودة في الدرجات الأدنى من المقياس.

والتوصية ITU-R BS.1284 لا تتضمن إلا طرائق مخصصة للمدى السمعي عالي الجودة أو لا تعطي أي درجات مطلقة للجودة السمعية.

وتركز توصيات أخرى مثل التوصية ITU-T P.800، وITU-T P.810، وITU-T P.830 على التقييم الشخصي للإشارات الكلامية في بيئة هاتفية. وأجرى فريق مشروع اتحاد الإذاعات الأوروبية B/AIM تجارب باستعمال مادة سمعية تقليدية مثل تلك المستعملة في البيئة الإذاعية باستعمال طرائق قطاع تقييس الاتصالات. ولا تلبي أي من هذه الطرائق متطلبات المقياس المطلق، والمقارنة بإشارة مرجعية، وفواصل ثقة إحصائية صغيرة مع عدد معقول من المقيّمين في نفس الوقت. ولذلك، لا يمكن إجراء تقييم للإشارات السمعية في بيئة إذاعية بالشكل الصحيح باستعمال إحدى هذه الطرائق.

وتهدف طريقة الاختبار المراجعة الوارد وصفها في هذه التوصية إلى تقديم إجراء موثوق ويمكن تكراره للأنظمة التي تقع جودتها السمعية عادةً في النصف الأدنى من مقياس الانحطاط المستعمل في التوصية ITU-R BS.1116 [2؛ و4؛ و3]. وفي طريقة الاختبار MUSHRA، تُستعمل إشارة مرجعية عالية الجودة ويتوقع أن تحدث الأنظمة التي يجري اختبارها انحطاطاً كبيراً. ويتعين استعمال طريقة MUSHRA في تقييم الأنظمة السمعية المتوسطة. وإذا ما استعملت طريقة MUSHRA مع المحتوى الملائم، فإن الوضع المثالي هو أن تتراوح درجات المستمعين بين 20 و80 نقطة MUSHRA. وإذا وقعت درجات معظم ظروف الاختبار في المدى 80-100 فقد يكون صحيحاً أن نتائج الاختبار غير صالحة.

والأسباب المحتملة للدرجات المضغوطة هي: الاستعانة بمقيّمين من غير ذوي الخبرة، أو استعمال محتوى غير حرج، أو اختيار اختبار غير ملائم لخوارزميات التشفير المستعملة في الاختبار.

3 التصميم التجريبي

تُستعمل أنواع مختلفة عديدة من استراتيجيات البحث في جمع معلومات موثوقة في مجال من مجالات الاهتمام العلمي. وستستعمل في التقييم الشخصي للانحطاط في الأنظمة السمعية أكثر الطرائق التجريبية التزاماً بالشكلية. وتتميز التجارب الشخصية أولاً بالتحكم والمعالجة الفعليين للظروف التجريبية، وثانياً بجمع وتحليل البيانات الإحصائية من المستمعين. وثمة حاجة إلى التصميم والتخطيط التجريبيين الدقيقين لضمان التقليل إلى أدنى حد العوامل التي يصعب التحكم فيها والتي يمكن أن تؤدي إلى عدم وضوح النتائج. ومثالاً على ذلك، إذا كان التسلسل الفعلي للمواد السمعية متماثلاً في نظر جميع المقيّمين في اختبار استماع، فإن المرء لا يستطيع التأكد مما إذا كانت الأحكام التي يصدرها المقيّمون ترجع إلى التسلسل وليس إلى المستويات المختلفة للانحطاط التي قُدّمت. وبناءً على ذلك، يجب ترتيب ظروف الاختبار بطريقة تكشف عن تأثيرات العوامل المستقلة، والعوامل التي تكشف عن هذه التأثيرات فقط.

وفي الحالات التي يمكن فيها توقع توزيع مظاهر الانحطاط والخصائص الأخرى المحتملة بطريقة متجانسة في جميع مراحل اختبار الاستماع، يمكن تطبيق شكل حقيقي من العشوائية في تقديم ظروف الاختبار. وحيث يُتوقع حدوث شيء من عدم التجانس يجب أن يوضع ذلك في الاعتبار عند تقديم ظروف الاختبار. فحيث يختلف مستوى صعوبة المادة المراد تقييمها، مثلاً، يجب توزيع نظام تقديم الحوافز بشكل عشوائي، داخل الجلسات وفيما بينها.

ويجب تصميم اختبارات الاستماع بحيث لا يُحَمّل المقيّمون بأكثر من طاقتهم إلى درجة تنتقص من دقة الأحكام. وباستثناء الحالات التي تكون العلاقة فيها بين الصوت والبصر مهمة، يُفضّل أن يجري تقييم الأنظمة السمعية غير مصحوبة بالصور. ومن الاعتبارات الرئيسية إدخال ظروف التحكم الملائمة. وتشمل ظروف التحكم، عادةً، تقديم مواد سمعية خالية من مظاهر الانحطاط يجري تقديمها للمقيّمين بطرائق لا يمكن التنبؤ بها. فالاختلافات بين أحكام هذه الحوافز الضابطة وأحكام الحوافز التي يحتمل أن تكون مقترنة بمظاهر الانحطاط هي التي تسمح بالخلوص إلى أن الدرجات هي تقييمات فعلية لمظاهر الانحطاط.

ويرد في موضع لاحق وصف لبعض هذه الاعتبارات. وينبغي أن يكون مفهوماً أن موضوعات التصميم التجريبي، والتنفيذ التجريبي، والتحليل الإحصائي تتسم بالتعقيد، وأنه لا يمكن إيراد كل التفاصيل في توصية كهذه. ويوصَى باستشارة متخصصين ذوي خبرة في التصميم التجريبي والإحصاء أو الاستعانة بهم في بداية التخطيط لاختبار الاستماع.

ولتمكين التحليل الكفء للبيانات ونقلها بين المختبرات، يجب الإبلاغ عن التصميم التجريبي. وينبغي تعريف كل من المتغيرات التابعة والمستقلة بالتفصيل. ويتم تعريف عدد المتغيرات المستقلة مع المستويات المرتبطة بها.

4 اختيار المقيّمين

ينبغي أن تأتي البيانات المستمدة من الاختبارات المقيّمة للانحطاط الضعيف في الأنظمة السمعية، كما يرد في التوصية ITU-R BS.1116، من مقيّمين ذوي خبرة في اكتشاف مظاهر الانحطاط الضعيف هذه. وكلما ارتفعت الجودة التي تصل إليها الأنظمة المراد اختبارها، كلما زادت أهمية اختيار مقيّمين ذوي خبرة.

1.4 معايير اختيار المقيّمين

في حين لا تهدف طريقة MUSHRA إلى تقييم الانحطاط الضعيف، فإنه يوصَى رغم ذلك بالاستعانة بمقيّمين ذوي خبرة لضمان جودة بيانات الاختبار التي يتم جمعها. وينبغي أن تتوفر لأولئك المستمعين الخبرة في الاستماع إلى الصوت بطريقة ناقدة. فهؤلاء المستمعون سيقدمون نتائج أكثر موثوقية وبسرعة أكبر من المستمعين غير ذوي الخبرة. ومن المهم أيضاً الإشارة إلى أن معظم المستمعين غير ذوي الخبرة يكونون عادة أكثر حساسية للأنواع المختلفة من الأصوات المصطنعة بعد التعرّض المتكرر لها. ويجري اختيار المقيّم ذي الخبرة بناءً على قدرته على إجراء اختبار الاستماع. ويجب تحديد جودة وكمية هذه القدرة بناءً على مهارات الموثوقية والتمييز لدى المقيّمين في الاختبار على أساس تكرار الاختبارات، على النحو الموضّح أدناه:

- **التمييز**: مقياس للقدرة على إدراك الاختلافات بين مواد الاختبار.
- **الموثوقية**: مقياس لتقارب عمليات التقييم المتكررة لنفس المادة من مواد الاختبار.

ولا يجب أن يُدرج في التحليل النهائي للبيانات إلا المقيّمون المصَنّفون كمقيّمين ذوي خبرة في أي اختبار معيّن. ويتوفر عدد من التقنيات اللازمة لأداء هذا التحليل للمقيّمين. ولمزيد من المعلومات يمكن الاطلاع على التقرير ITU-R BS.2300.[1] وتقوم هذه التقنيات على أساس واحدة على الأقل من عمليات التقييم المتكررة التي يقوم بها المقيّم وتتيح تحديد جودة وكمية خبرة المقيّم في التجربة الواحدة. ويجب تطبيق هذه الطرائق إما كفرز مسبق للمقيّمين في نطاق تجربة رائدة أو يفضّل أن تطبّق كفرز مسبق وجزء من الاختبار الرئيسي. وترتبط التجربة الرائدة بسلسلة من التجارب وتشمل مجموعة ممثلة من عينات الاختبار المراد تقييمها في التجربة

[1] من أمثلة تنفيذ هذه التقنية طريقة مقياس الخبرة (eGuage) الوارد وصفها في التقرير ITU-R BS.2300-0. وهي متاحة على الموقع الإلكتروني http://www.itu.int/oth/R0A07000036.

الرئيسية. ولتقييم خبرة المستمع، يجب أن تشمل التجربة الرائدة مجموعة فرعية ذات صلة من حوافز الاختبار تمثل المدى الكامل للحوافز والأصوات المصطنعة المراد تقييمها أثناء التجربة (التجارب) الرئيسية الفعلية.

وينبغي أن ينقل التمثيل البياني للتحليل معلومات عن موثوقية المستمعين مقابل قدرتهم على التمييز.

1.1.4 الفرز المسبق للمقيّمين

ينبغي أن تتكون مجموعة الاستماع من مستمعين ذوي خبرة، وبعبارة أخرى، من أشخاص يفهمون الطريقة الموصوفة للتقييم الشخصي للجودة والذين جرى تدريبهم عليها كما ينبغي. وينبغي أن تتوفر لهؤلاء المستمعين:

- الخبرة في الاستماع إلى الصوت بطريقة ناقدة؛
- قدرة سمع طبيعية (يجب استعمال المعيار 389 للمنظمة الدولية لتوحيد المقاييس كمبدأ توجيهي).

وينبغي استعمال إجراء التدريب كأداة للفرز المسبق. ولا يُدرَج عند تحليل البيانات إلا *المستمعون المصَنّفون كمقيّمين ذوي خبرة* إما في التجربة الرائدة أو في التجربة الرئيسية.

ويستعمل إدراج تكرارات الحوافز لتوفير طريقة لتقييم موثوقية المستمعين.

والحجة الرئيسية لإدخال تقنية للفرز المسبق هي زيادة كفاءة اختبار الاستماع. غير أنه يتعين موازنة ذلك بالمخاطرة المتمثلة في الحد من أهمية النتيجة على نحو مبالغ فيه.

2.1.4 الفرز اللاحق للمقيّمين

تستبعد طريقة الفرز اللاحق المقيّمين الذين يعطون درجة مرتفعة جداً لإشارة مرتكزات تتضمن انحطاطاً كبيراً، وأولئك الذين كثيراً ما يقيّمون المرجع المحجوب كما لو كانت تنطوي على انحطاط كبير، على النحو المحدد في المقاييس التالية:

- ينبغي استبعاد المقيّم من الردود المجمّعة إذا قيّم حالة المرجع المحجوب لأكثر من 15 في المائة من مواد الاختبار بأقل من 90 درجة؛
- ينبغي استبعاد المقيّم من الردود المجمّعة إذا قيّم مرتكز متوسط المدى لأكثر من 15 في المائة من مواد الاختبار بأكثر من 90 درجة. وإذا قيّم أكثر من 25 في المائة من المقيّمين مرتكز متوسط المدى بأكثر من 90 درجة، فإن ذلك قد يشير إلى أن مادة الاختبار لم تُخفض درجتها بشكل كبير نتيجة لمعالجة المرتكز. وفي هذه الحالة، لا ينبغي استبعاد المقيّمين على أساس الدرجات المعطاة لهذه المادة.

ويمكن تنفيذ هذه المرحلة المبدئية قبل أن ينتهي جميع المقيّمين من اختباراتهم إذا طُلب ذلك (بما يتيح لمختبر الاختبارات أن يقيّم ما إذا كان لديه العدد الكافي من المقيّمين قبل انتهاء الاختبارات).

وقد يكون من المفيد دراسة البيانات لتحديد نقاط البيانات النائية الخاطئة لإخضاعها لمزيد من التحليل. ومن الطرائق المناسبة استعمال مقارنة الدرجات الفردية بالمدى الرُبيعي لجميع الدرجات المعطاة لحالة اختبار معينة *j*، والتسلسل السمعي *k*.

وينبغي أن يُحسب المتوسط $\hat{x}$ والرُبيعات Q على النحو التالي:

$$\hat{x} := Q_2(x_{jk}) = \text{median}(x) := \begin{cases} x_{jk\frac{n+1}{2}}, n \text{ odd} \\ \frac{1}{2}\left(x_{jk\frac{n}{2}} + x_{jk\frac{n}{2}+1}\right), n \text{ even} \end{cases}$$

وتُرتب x حسب الحجم من الأصغر إلى الأكبر

$$Q_1(x_{jk}) = \begin{cases} \text{median}\left(x_{jk1}, \ldots, x_{jk\frac{n+1}{2}}\right), n \text{ odd} \\ \text{median}\left(x_{jk1}, \ldots, x_{jk\frac{n}{2}}\right), n \text{ even} \end{cases},$$

$$Q_3(x_{jk}) = \begin{cases} \text{median}\left(x_{jk1}, \dots, x_{jk\frac{n+1}{2}}\right), n \text{ odd} \\ \text{median}\left(x_{jk\frac{n}{2}+1}, \dots, x_{jkn}\right), n \text{ even} \end{cases}.$$

ويُحسب المدى الرُبيعي بوصفه $IQR(x) \coloneqq Q_3(x) - Q_1(x)$.

وفي هذا السياق، تنتمي القيم الشاذة إلى المجموعة $O(x_{jk})$:

$$O(x_{jk}) \coloneqq \{x_{jk} | x_{jk} > Q_3(x_{jk}) + 1.5 \cdot IQR(x_{jk})\} \cup \{x_{jk} | x_{jk} < Q_1(x_{jk}) - 1.5 \cdot IQR(x_{jk})\}.$$

وإذا أعطي أحد أفراد عينة الاختبار درجة x لحافز معين وكان النظام الذي يجري اختباره عنصراً من $O(x)$، ينبغي فحص أسباب إعطاء هذه الدرجة. وقد يكشف فحص تسجيل جلسة الاختبار عن مشاكل تقنية في الأجهزة، أو عن خطأ بشري. كما يمكن أن يكشف سؤال المقيّم عمّا إذا كانت الدرجة المعطاة تمثل رأيه الشخصي حقاً. وإذا تبين أن السبب في وجود نقطة البيانات الشاذة هو خطأ، يمكن حذفها من مجموعة البيانات قبل التحليل النهائي، مع ذكر سبب حذفها في تقرير الاختبار.

وقد يوضح تطبيق طريقة الفرز اللاحق الاتجاهات في نتيجة الاختبار. غير أنه إذا وُضع في الاعتبار تباين حساسيات المقيّمين تجاه الأصوات المصطنعة، فينبغي توخي الحذر. وعن طريق زيادة حجم مجموعة الاستماع يمكن الحد من آثار أي درجات فردية للمقيّمين.

2.4 حجم مجموعة الاستماع

يمكن تحديد الحجم المناسب لمجموعة الاستماع إذا أمكن تقدير تباين الدرجات التي يعطيها المقيّمون المختلفون وعُرف ثبات التجربة المطلوب.

وحيث تخضع ظروف اختبار الاستماع للتحكم الصارم من الجانبين التقني والسلوكي، أظهرت التجربة أن البيانات الواردة من عدد لا يزيد عن عشرين مقيّماً غالباً ما تكفي للخلوص إلى استنتاجات ملائمة من الاختبار. وإذا أمكن إجراء تحليل مع تقدم الاختبار، فلن تكون هناك حاجة إلى معالجة مزيد من المقيّمين بعد الوصول إلى مستوى مناسب من الدلالة الإحصائية للخلوص إلى استنتاجات ملائمة من الاختبار.

وإذا لم يتيسر التحكم بشكل صارم في التجربة لأي سبب من الأسباب، قد تكون هناك حاجة إلى عدد أكبر من المقيّمين للوصول إلى الثبات المطلوب.

ولا يتعلق حجم مجموعة الاستماع بالثبات المطلوب فقط. فالنتيجة التي يتم الوصول إليها من نوع التجربة الذي تُعنى به هذه التوصية لا تصلح من حيث المبدأ إلا لمجموعة المستمعين ذوي الخبرة المشتركين فعلاً في الاختبار. وبالتالي، يمكن القول إنه بزيادة حجم مجموعة الاستماع ستنطبق النتيجة على مجموعة أكثر عمومية من المستمعين ذوي الخبرة، ويمكن لذلك اعتبارها أحياناً أكثر إقناعاً. وقد تكون هناك حاجة أيضاً إلى زيادة حجم مجموعة الاستماع للسماح باحتمال تباين المقيّمين في حساسيتهم تجاه الأصوات المصطنعة المختلفة.

5 طريقة الاختبار

تَستعمل طريقة MUSHRA مادة البرامج غير المعالجة الأصلية بنفس عرض النطاق الكامل المستعمل للإشارة المرجعية (المستعملة أيضاً كمرجع محجوب) وكذلك عدد من المرتكزات المحجوبة الإلزامية.

ويمكن استعمال مرتكزات محجوبة إضافية، ويفضّل أن تكون من بين تلك التي تمثل موضوع توصيات أخرى ذات صلة لقطاع الاتصالات الراديوية. وبالنظر إلى أن خواص المرتكزات يمكن أن يكون لها تأثير كبير على نتائج الاختبار، ينبغي أن يضع تصميم المرتكز غير القياسي في الاعتبار سلوك المرتكزات المثلى الوارد وصفها في المرفق 5. وينبغي وصف طبيعة أي مرتكزات غير قياسية تستعمل في الاختبار بالتفصيل في تقرير الاختبار.

1.5 وصف إشارات الاختبار

يوصى بأن يكون أقصى طول للتسلسلات 10 s تقريباً، ويفضل ألا يتجاوز 12 s. والهدف من هذا هو تجنب إجهاد المستمعين، وزيادة الدقة والثبات في أجوبة المستمعين، وتقصير المدة الكلية لاختبار الاستماع. وهذه المدة ضرورية أيضاً لتيسير اتساق المحتوى طوال مدة الإشارة والتي ينبغي أن تزيد الاتساق في أجوبة المستمعين. وبالإضافة إلى ذلك، من شأن تقصير المدة أيضاً أن يتيح للمستمعين مقارنة قدر أكبر من إشارات الاختبار المتواصلة.

وإذا كانت الإشارات أطول مما ينبغي، فإن أجوبة المستمعين توجهها تأثيرات أسبقية وحداثة إشارات الاختبار أو مناطق عروية منعزلة يمكن أن تتباين كثيراً في السمات الطيفية والزمنية خلال مدة إشارة الاختبار. ويهدف تقصير مدة إشارات الاختبار إلى تقليل هذا التباين. غير أن هذا القيد قد لا يكون ملائماً في بعض الظروف. ويمكن أن يكون من الأمثلة على ذلك اختبار يتضمن مساراً متحركاً بطيئاً وطويلاً لصوت ما. وفي هذه الظروف المقيدة التي يتقرر فيها وجوب استعمال حافز أطول أمداً، من الضروري توثيق مبررات هذا الشرط المتعلق بزيادة المدة في تقرير الاختبار النهائي.

وتتألف مجموعة الإشارات المعالَجة من كل الإشارات التي يجري اختبارها وعلى الأقل "مرتكزين" إضافيين. والمرتكز القياسي هو نسخة مرشّحة بتمرير منخفض من الإشارة الأصلية بتردد قطع قدره 3,5 kHz؛ وتردد قطع مرتكز متوسط الجودة قدره 7 kHz.

وتُطابق عروض نطاقات المرتكزات التوصيات في دارات التحكم (3,5 kHz) المستعملة لأغراض الإشراف والتنسيق في الإذاعة ودارات التعليق (7 kHz) والدارات العَرَضية (10 kHz)، وفقاً للتوصيات ITU-T G.711، وITU-T G.712، وITU-T G.722، وITU-T J.721، على التوالي.

وينبغي أن تكون خصائص مرشاح التمرير المنخفض الذي تبلغ قدرته 3,5 kHz كما يلي:

fc = 3,5 kHz

الحد الأقصى لتموّج نطاق التمرير = ± 0,1 dB

الحد الأدنى للتوهين عند 4 kHz = 25 dB

الحد الأدنى للتوهين عند 4,5 kHz = 50 dB.

والهدف من المرتكزات الإضافية هو إعطاء مؤشر إلى كيف تُقارن الأنظمة التي يجري اختبارها بسويات الجودة السمعية المعروفة جيداً ولا ينبغي استعمالها لإعادة تحديد قياسات النتائج بين الاختبارات المختلفة.

2.5 مرحلة التدريب

من أجل تحقيق نتائج موثوقة، يلزم تدريب المقيّمين في جلسات تدريب خاصة قبل الاختبار. وقد تبيّن أن هذا التدريب مهم لتحقيق نتائج موثوقة. ويجب أن يُعرّض التدريب الأفراد المشاركين في الاختبار للمدى الكامل للانحطاط وطبيعته وكل إشارات الاختبار التي سيتعرّضون لها أثناء الاختبار. ويمكن تحقيق ذلك باستعمال عدة طرائق مثل نظام بسيط لإعادة تشغيل الشريط أو نظام تفاعلي يتم التحكم فيه بالحاسوب. وترد تعليمات بهذا الشأن في المرفق 1. كما ينبغي استعمال التدريب لضمان دراية المقيّمين ببنية الاختبار الشخصي (مثل برمجية الاختبار).

3.5 عرض الحوافز

يعتبر اختبار MUSHRA طريقة اختبار بحجب مزدوج متعدد الحوافز بمرجع محجوب ومرتكزات محجوبة، في حين تستعمل التوصية ITU-R BS.1116 طريقة اختبار "الحجب المزدوج الثلاثي الحوافز مع المرجع المحجوب". وثمة شعور بأن نهج MUSHRA أنسب لتقييم الانحطاط المتوسط والكبير [4].

وفي الاختبار الذي يتضمن مظاهر انحطاط ضعيف، تتمثل المهمة الصعبة للفرد المشارك في الاختبار في اكتشاف أي أصوات مصطنعة قد تكون موجودة في الإشارة. وفي هذه الحالة تكون هناك حاجة إلى إشارة مرجعية محجوبة في الاختبار لتساعد القائم

بتنفيذ التجربة في تقييم قدرة المقيّم على اكتشاف هذه الأصوات المصطنعة بنجاح. وعلى العكس من ذلك، لن يجد الفرد المشارك في الاختبار صعوبة في اكتشاف الأصوات المصطنعة في الاختبار الذي يتضمن مظاهر انحطاط متوسطة وكبيرة ولذلك لا تكون هناك حاجة إلى إشارة مرجعية محجوبة لهذا الغرض. ولكن تبدأ المشاكل عندما يتعين على الفرد المشارك في الاختبار وضع درجة لمظاهر الإزعاج النسبية للأصوات المصطنعة المختلفة. وهنا يتعين على الفرد المشارك في الاختبار الموازنة بين تفضيله لنوع معين من الأصوات المصطنعة وتفضيله لنوع آخر منها.

ويطرح استعمال المرجع عالي الجودة مشكلة مثيرة. فنظراً إلى أن المنهجية الجديدة يتعين استعمالها لتقييم الانحطاط المتوسط والكبير، فإن من المتوقع أن يكون الاختلاف الإدراكي بين الإشارة المرجعية ومواد الاختبار كبيراً نسبياً. وعلى العكس من ذلك، فإن الاختلافات الإدراكية بين مواد الاختبار الخاصة بالأنظمة المختلفة قد تكون صغيرة. ونتيجة لذلك، فإذا استعملت طريقة اختبار تعتمد على المحاولات المتعددة (مثل تلك المستعملة في التوصية ITU-R BS.1116)، قد يجد المقيّمون صعوبة كبيرة في التمييز بدقة بين الإشارات المختلفة التي تتضمن انحطاطاً. وعلى سبيل المثال، قد يتفق المقيّمون في اختبار مقارنة زوجية مباشرة على أن النظام ألف أفضل من النظام باء. غير أنه في حالة يقارَن فيها كل نظام بالإشارة المرجعية فقط (أي أن النظام ألف والنظام باء لا يقارَنان كل منهما بالآخر بشكل مباشر)، قد تُفقد الاختلافات بين النظامين.

وللتغلب على هذه المشكلة، يمكن للفرد المشارك في الاختبار، في طريقة اختبار MUSHRA، أن يتنقّل بإرادته بين الإشارة المرجعية وأي من الأنظمة التي يجري اختبارها، عادةً باستعمال نظام إعادة التشغيل القائم على الحاسوب، وإن كان يُمكن استعمال آليات أخرى تستعمل آلات الأقراص المدمَجة أو الشرائط المتعددة. فتُعرَض على الفرد المشارك في الاختبار سلسلة من التجارب. وتُعرَض على الفرد في كل تجربة النسخة المرجعية والمرتكز المنخفض والمتوسط، وجميع نُسخ إشارة الاختبار المعالَجة بالأنظمة التي يجري اختبارها. فإذا كان النظام يحتوي، مثلاً، على 8 أنظمة سمعية، يُسمَح للفرد بالتنقل في نفس الوقت تقريباً بين إشارات الاختبار الإحدى عشرة والمرجع المفتوح (مرجع واحد + 8 أنظمة اختبار + مرجع محجوب واحد + مرتكز منخفض محجوب واحد + مرتكز متوسط محجوب واحد).

ونظراً إلى أنه يمكن للفرد أن يقارن مباشرةً الإشارات ذات الانحطاط، فإن هذه الطريقة توفر مزايا اختبار المقارنة الزوجية الكاملة حيث يستطيع الفرد أن يكتشف بسهولة أكبر الاختلافات بين الإشارات التي تنطوي على انحطاط وتحديد درجاتها تبعاً لذلك. وتتيح هذه السمة درجة عالية من الوضوح في الدرجات المعطاة للأنظمة. غير أنه من المهم ملاحظة أن المقيّمين سيستخلصون الدرجات التي يعطونها لنظام ما بمقارنة هذا النظام بالإشارة المرجعية وكذلك بالإشارات الأخرى في كل تجربة.

ويوصَى بألا تحتوي أي تجربة على أكثر من 12 إشارة (مثلاً، 9 أنظمة يجري اختبارها، ومرتكز منخفض محجوب واحد، ومرتكز متوسط محجوب واحد، ومرجع محجوب واحد).

وفي الحالة النادرة التي يتعين فيها مقارنة عدد كبير من الإشارات، يمكن أن تكون هناك حاجة إلى استعمال تصميم محجوب للتجربة، ويجب الإبلاغ عن ذلك بالتفصيل.

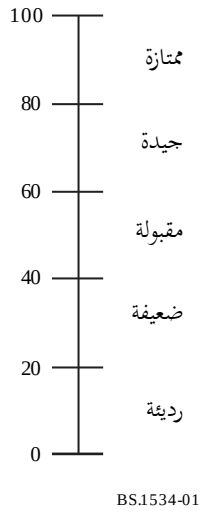
وفي اختبارات التوصية ITU-R BS.1116، يميل المقيّمون إلى الخوض في تجربة معينة بالبدء بعملية اكتشاف، تليها عملية تحديد الدرجات. وتُبيّن الخبرة المستمدة من إجراء الاختبارات وفقاً لطريقة MUSHRA أن المقيّمين يميلون إلى بدء الجلسة بتقدير تقريبي للجودة. وتلي ذلك عملية تقييم أو ترتيب. وبعد ذلك يؤدي الفرد المشارك في الاختبار عملية تحديد الدرجات. ونظراً إلى أن عملية الترتيب تتم بطريقة مباشرة، فإن نتائج الجودة السمعية المتوسطة يحتمل أن تكون أكثر اتساقاً وموثوقية مما لو استعملت الطريقة الواردة في التوصية ITU-R BS.1116. وبالإضافة إلى ذلك، فإن الحد الأدنى لمدة العروة هو 500 ms وينبغي تطبيق خبو داخل وخبو خارج لغلاف جيب التمام المرفوع بمقدار 5 ms على جميع المحتويات التي تكون في شكل عروة. وينبغي أن تتضمن جميع عمليات التبديل بين محتوى أنظمة الاختبار خبو داخل مقداره 5 ms وخبو خارج مقداره 5 ms لغلاف جيب التمام المرفوع. ولا ينبغي في أي وقت من الأوقات أثناء أي اختبار استعمال الخبو التبادلي عند التنقل بين أنظمة الاختبار. وتهدف هذه التعديلات إلى الحد من استعمال التغيرات في التلون الطيفي أثناء المقارنات العارضة المفاجئة لتحديد وتقييم الإشارات التي يجري اختبارها.

4.5 عملية تحديد الدرجات

على المقيّمين تحديد درجة للحوافز وفقاً لمقياس الجودة المتواصل (CQS). ويتكون هذا المقياس من مقاييس بيانية متشابهة (عادةً ما يكون طولها 10 سم أو أكثر) مقسمةً إلى خمسة فواصل متساوية تُطلق عليها صفات على النحو الوارد في الشكل 1 من أعلى إلى أسفل.

ويستعمل هذا المقياس أيضاً لتقييم جودة الصورة (التوصية ITU-R BT.500 – منهجية التقييم الشخصي لجودة الصورة التلفزيونية).

الشكل 1



ويسجل المستمع تقييمه للجودة بشكل مناسب، مثلاً، باستعمال مساطر زلاقة على شاشة إلكترونية (انظر الشكل 2)، أو باستعمال مقياس القلم والورقة. وعند استعمال بنية مماثلة لتلك الموضحة في الشكل 2، ينبغي تقييد تصرفات الفرد، بحيث لا يستطيع أن يعدل إلا الدرجة المعطاة للمادة التي يستمع إليها في تلك اللحظة. ويمكن الرجوع إلى المرفق 2 للحصول على بعض التوجيهات المتعلقة بتصميم السطح البيني. ويطلب من المقيّم أن يقيّم جودة الحوافز وفقاً لمقياس الجودة المتواصل ذي الفواصل الخمسة.

الشكل 2

مثال على استعمال شاشة الحاسوب في اختبار MUSHRA

BS.1534-02

ومقارنةً بالتوصية ITU-R BS.1116، فإن طريقة MUSHRA تتميز بأنها تقدم عدداً كبيراً من الحوافز في نفس الوقت بحيث يستطيع الفرد تنفيذ أي مقارنة بينها بشكل مباشر. ويمكن أن يُخفّض بدرجة ملموسة الوقت المستغرق في أداء الاختبار بطريقة MUSHRA مقارنةً باستعمال الطريقة الواردة في التوصية ITU-R BS.1116.

5.5 تسجيل جلسات الاختبار

في حالة ملاحظة شيء شاذ أثناء معالجة الدرجات المعطاة، من المفيد جداً وجود سجل بالأصوات التي نتجت عنها الدرجات. ومن الوسائل البسيطة نسبياً لتحقيق ذلك إعداد تسجيلات فيديوية وسمعية للاختبار كله. وفي حالة العثور على درجة شاذة في مجموعة من النتائج، يمكن فحص تسجيل الشريط لمحاولة التأكد مما إذا كان السبب خطأً بشرياً أو خللاً وظيفياً في الأجهزة.

6 النعوت

ترد فيما يلي النعوت المحددة لتقييمات الأصوات غير المجسمة والأصوات المجسمة والقنوات المتعددة. ويفضل تقييم نعت "الجودة السمعية الأساسية" في كل حالة. وقد يرغب القائمون بتنفيذ التجربة في تعريف وتقييم نعوت أخرى.

وينبغي تحديد درجات نعت واحد فقط أثناء التجربة. وعندما يُطلب من المقيّمين تقييم أكثر من نعت واحد في كل تجربة، فإن ذلك قد يحملهم فوق طاقتهم أو يسبب لهم ارتباكاً أو يؤدي إلى الاثنين معاً، وذلك عن طريق محاولة الإجابة على أسئلة متعددة عن حافز معين. وقد يؤدي ذلك إلى تحديد الدرجات بشكل غير موثوق لكل الأسئلة. وإذا كان المطلوب تقييم خواص متعددة للمادة السمعية بشكل مستقل، فإنه يوصَى بتقييم الجودة السمعية الأساسية أولاً.

1.6 النظام غير المجسم

الجودة السمعية الأساسية: يستعمل هذا النعت الشامل الوحيد لتقييم أي وكل من الاختلافات التي يتم اكتشافها بين المرجع والشيء.

2.6 النظام المجسم

الجودة السمعية الأساسية: يستعمل هذا النعت الشامل الوحيد لتقييم أي وكل من الاختلافات التي يتم اكتشافها بين المرجع والشيء.

وقد يكون النعت الإضافي التالي ذا أهمية:

جودة الصورة الصوتية المجسمة: يتعلق هذ النعت بالاختلافات بين المرجع والشيء من حيث مواقع الصور الصوتية والإحساس بعمق وواقعية الحدث السمعي. ورغم أن بعض الدراسات قد أظهرت أن جودة الصورة الصوتية المجسمة يمكن أن يصيبها الانحطاط، فإنه لم تجر بحوث كافية حتى الآن لتوضيح ما إذا كانت هناك حاجة إلى تقييم لجودة الصورة الصوتية المجسمة منفصل عن الجودة السمعية الأساسية.

الملاحظة 1– حتى عام 1993، استعملت معظم الدراسات المتعلقة بالتقييم الشخصي للانحطاط الضعيف للأنظمة المجسمة نعت الجودة السمعية الأساسية حصراً. وبالتالي أُدمِج نعت جودة الصورة الصوتية المجسمة بشكل ضمني أو صريح في الجودة السمعية الأساسية بوصفه نعتاً شاملاً في هذه الدراسات.

3.6 النظام المتعدد القنوات

الجودة السمعية الأساسية: يستعمل هذا النعت الشامل الوحيد لتقييم أي وكل من الاختلافات التي يتم اكتشافها بين المرجع والشيء.

وقد تكون النعوت التالية ذات أهمية:

جودة الصورة الأمامية: يتعلق هذا النعت بتحديد موقع مصادر الصوت الأمامية. ويشمل جودة الصورة الصوتية المجسمة وخسائر الاستبانة.

نعت انطباع الوسط المحيط: يتعلق هذا النعت بالانطباع المكاني، أو الجو العام، أو تأثيرات الوسط المحيط الاتجاهية الخاصة.

7 مادة الاختبار

يجب أن تُستَعمل مادة حرجة تمثل البرنامج الإذاعي النموذجي للتطبيق المطلوب للكشف عن الاختلافات بين الأنظمة التي يجري اختبارها. وتعتبر المادة حرجة إذا كانت تركز على الأنظمة التي يجري اختبارها. ولا توجد مواد برامج "مناسبة" لجميع الحالات ويمكن استعمالها لتقييم جميع الأنظمة وفي جميع الظروف. وبناءً على ذلك، يجب السعي إلى العثور على مواد البرامج الحرجة بشكل صريح لكل نظام يراد اختباره في كل تجربة. وعادة ما يستهلك البحث عن المواد المناسبة الكثير من الوقت؛ غير أنه ما لم يتم العثور على المواد الحرجة حقاً لكل نظام، فلن تنجح التجارب في الكشف عن الاختلافات بين الأنظمة وستكون غير قاطعة. وينبغي أن تقوم مجموعة صغيرة من المستمعين ذوي الخبرة باختيار مواد الاختبار من بين مجموعة أكبر من المواد المرشحة المحتملة. ويجب أن تشمل عملية الاختيار هذه جميع أنظمة الاختبار وتوثيقها والإبلاغ عنها في ملخص الاختبار.

ويجب أن يثبت تجريبياً وإحصائياً أن أي إخفاق في العثور على اختلافات بين الأنظمة لا يرجع إلى انعدام الحساسية التجريبية بسبب الاختيارات السيئة للمواد السمعية، أو أي جوانب ضعف أخرى في التجربة. وخلاف ذلك، لا يمكن قبول النتيجة "الصفرية" كنتيجة سليمة.

وعند البحث عن المواد الحرجة، يُسمح بأي حافز يمكن اعتباره مادة إذاعية محتملة. ولا ينبغي أن يشمل ذلك أي إشارات تركيبية مصممة لكسر نظام محدد عن قصد. ولا ينبغي أن يكون المحتوى الفني أو الفكري لأي تتابع برامجي شديد الجاذبية أو شديد التنفير أو باعثاً على الملل إلى درجة يتشتت معها الفرد تركيزه فلا يستطيع اكتشاف مظاهر الانحطاط. وينبغي أن يوضع في الاعتبار التكرار المتوقع لوقوع كل نوع من أنواع المواد البرنامجية في عمليات الإذاعة الفعلية. غير أنه ينبغي أن يكون مفهوماً أن طبيعة المواد المذاعة قد تتغير مع مرور الوقت مع حدوث تغيرات مستقبلية في الأساليب والتفضيلات الموسيقية.

وعند اختيار المواد البرنامجية، من المهم تعريف النعوت المراد تقييمها بدقة. وستلقى مسؤولية اختيار المواد على عاتق فريق من الأفراد المهرة ممن لديهم معرفة أساسية بمظاهر الانحطاط التي يمكن توقعها. وستعتمد نقطة انطلاقهم على مجموعة واسعة جداً من المواد. ويمكن زيادة المجموعة بتسجيلات مخصصة.

ولإعداد أشرطة اختبار المقارنة الشخصية، يلزم تعديل درجة ارتفاع صوت كل مقتطف بطريقة شخصية من قِبَل فريق الأفراد المهرة قبل تسجيلها على متوسط الاختبار. وسوف يتيح ذلك استعمال متوسط الاختبار لاحقاً في بيئة كسب ثابتة لجميع المواد البرنامجية.

ولذلك، فإن فريق الأفراد المهرة سوف يجتمع، فيما يتعلق بجميع تتابعات الاختبار، ويصل إلى توافق في الآراء بشأن مستويات الصوت النسبية لمقتطفات الاختبار الفردية. وبالإضافة إلى ذلك، ينبغي أن يصل الخبراء إلى توافق في الآراء حول مستوى ضغط الصوت المستنسخ المطلق للتتابع ككل مقارنة بمستوى التوحيد.

وينبغي إدراج رشقة نغمة (مثلاً 1 kHz، و300 ms، و–18 dBFS) على مستوى إشارة التراصف على رأس كل تسجيل حتى يمكن تعديل مستوى تراصفه الخارج لمستوى التراصف الداخل الذي تحتاج إليه قناة النسخ، وفقاً لتوصية اتحاد الإذاعات الأوروبية R.68 (انظر الفقرة 1.4.8 من التوصية ITU-R BS.1116). وتستعمل رشقة النغمة لأغراض التراصف فقط: ولا ينبغي إعادة تشغيلها أثناء الاختبار. وينبغي التحكم في إشارة البرنامج الصوتي بحيث لا تتجاوز اتساعات الذرى إلا نادراً اتساع ذروة أقصى إشارة مسموح بها على النحو الموضح في التوصية ITU-R BS.645 (موجة جيبية قدرها 9 dB فوق مستوى التراصف).

ويتباين العدد المقبول من المقتطفات التي يتعين إدراجها في الاختبار، ويجب أن يكون متساوياً لكل نظام يجري اختباره. والتقدير المعقول هو 1,5 مرة عدد الأنظمة التي يجري اختبارها مع مراعاة أن تكون القيمة الدنيا هي 5 مقتطفات. ونظراً إلى تعقد المهمة، يجب أن تكون الأنظمة التي يجري اختبارها متاحة للقائم بتنفيذ التجربة. ولا يمكن نجاح الاختيار إلا إذا تم تحديد جدول زمني ملائم. وبالإضافة إلى ذلك، ونظراً إلى استعمال معدل البتات لمتغير الوقت في الكودكات السمعية فإنه يوصَى بتشفير التتابعات الأطول واستعمال جزء من كل تتابع في اختبار الاستماع.

وسيستعمل خلط منخفض مرجعي لاختبار أداء نظام متعدد القنوات في ظروف إعادة التشغيل على قناتين. وعلى الرغم من أن استعمال خلط منخفض ثابت قد يعتبر مقيّداً في ظروف معينة، فإنه بدون شك أنسب خيار يمكن أن تستعمله الهيئات الإذاعية في الأجل الطويل. والمعادلتان المتعلقتان بالخلط المنخفض المرجعي (انظر التوصية ITU-R BS.775) هما:

$$L_0 = 1.00L + 0.71C + 0.71L_s$$

$$R_0 = 1.00R + 0.71C + 0.71R_s$$

وينبغي أن يستند الاختيار المسبق لمقتطفات الاختبار المناسبة لإجراء تقييم حرج لأداء خلط منخفض مرجعي على قناتين إلى نسخ مواد برنامجية خضعت إلى خلط منخفض على قناتين.

8 ظروف الاستماع

يرد في التوصية ITU-R BS.1116 تعريف لطرائق التقييم الشخصي للانحطاط الضعيف في الأنظمة السمعية، بما في ذلك الأنظمة الصوتية متعددة القنوات. ولتقييم الأنظمة السمعية ذات الجودة المتوسطة ينبغي استعمال ظروف الاستماع الواردة في الفقرتين 7 و8 من التوصية ITU-R BS.1116.

ويمكن استعمال سماعات الرأس أو مكبرات الصوت في الاختبار، ولا يُسمح باستعمال الاثنين في جلسة اختبار واحدة: ويجب على جميع المقيّمين استعمال نفس النوع من محوّل الطاقة.

وبالنسبة لإشارة قياس بجذر مربع تربيع توتر يساوي "مستوى إشارة تراصف" (0 dBµ0s وفقاً للتوصية ITU-R BS.645؛ و–18 dB تحت مستوى التقليم لتسجيل على شريط رقمي، وفقاً لتوصية اتحاد الإذاعات الأوروبية R.68) وتغذي بدورها دخل كل قناة من قنوات النسخ (أي مضخم القدرة ومكبر الصوت المرتبط به)، يجب تعديل كسب المضخم لينتج مستوى ضغط الصوت المرجعي (بترجيح -IEC/A، بطيء):

$$L_{ref} = 85 - 10 \log n \pm 0{,}25 \qquad \text{dBA}$$

حيث n هو عدد قنوات النسخ في التكوين الكلي.

ويُسمح للفرد المشارك في الاختبار بإجراء ضبط فردى لمستوى الاستماع أثناء الجلسة وينبغي أن يكون ذلك في حدود مدى ±4 dB مقارنةً بالمستوى المرجعي الوارد في التوصية ITU-R BS.1116. ويجب تحقيق التوازن بين مواد الاختبار في الاختبار الواحد عن طريق لجنة الاختيار بطريقة لا تجعل المقيّمين عادةً في حاجة إلى إجراء ضبط فردي لكل مادة.

ولا ينبغي السماح بضبط المستوى في المادة الواحدة.

9 التحليل الإحصائي

يتم تحويل تقييمات كل ظرف من ظروف الاختبار خطياً من قياسات الطول على كشف الدرجات إلى درجات مقيّسة في المدى 0 إلى 100 حيث يوافق 0 أدنى درجات المقياس (جودة رديئة). ثم تحسب الدرجات على النحو التالي:

ويمكن إجراء تحليل إحصائي معْلَمي أو غير معْلَمي، على أساس استيفاء الافتراضات الإحصائية (انظر الفقرة 3.3.9). وفيما يتعلق بالتوجيهات الخاصة بالتحليل الإحصائي المعْلَمي، انظر المرفق 4.

1.9 العرض التصويري للبيانات والتحليل الاستطلاعي للبيانات

ينبغي أن يبدأ التحليل الإحصائي دائماً بعرض تصويري للبيانات الخام. ويمكن أن يتضمن ذلك استعمال المدرج التكراري (histogram) مع منحنى مطابقة للتوزيع الطبيعي، أو مخططات الصندوق (boxplots)، أو المخططات الرُبيعية (quartile-quartile plots).

ويعطي تصوير البيانات بمخطط الصندوق إشارة إلى وجود وتأثير القيم الشاذة على الملخصات الوصفية للبيانات. ويجب إجراء هذا العرض التصويري لتحديد انتشار وانحراف الدرجات الفردية عن الدرجة المتوسطة لكل المقيّمين. وينبغي إجراء عرض تصويري باستعمال المدرج التكراري لتحديد وجود توزيع كامن متعدد المناويل. وإذا تم رؤية التوزيع المتعدد المناويل بوضوح في البيانات، يُنصح القائم بالتجربة بتحليل التوزيع بشكل مستقل.

ولتقييم درجة تعدد المناويل b، يمكن استعمال المعادلة التالية:

$$b = \frac{g^2 + 1}{k + \frac{3(n-1)^2}{(n-2)(n-3)}}$$

حيث:

n: هي حجم العينة

g: هي تخالف العينة المحدودة

k: هي التفرطح الزائد في نتائج اختبار الاستماع

ويقع هذا المعامل بين 0 و1. ويمكن تفسير القيم الأعلى (أكبر من 9/5) كإشارة إلى تعدد المناويل.

وبناءً على البحث بالنظر لهذه المخططات وb والافتراضات المتعلقة بالمجتمع الإحصائي للعينة الخاضعة للملاحظة، ينبغي تقرير ما إذا كان ينبغي للمرء أن يفترض أنه لاحظ توزيعا طبيعياً أم لا. وإذا كان من الواضح أن منحنى المطابقة متخالفاً أو المدرج التكراري يحتوي على كثير من القيم الشاذة أو المخطط الربيعي ليس خطاً مستقيماً على الإطلاق، فلا ينبغي للمرء أن يعتبر أن العينة موزعة توزيعاً طبيعياً. وسوف يؤدي حساب متوسط الدرجات المقيّسة لكل المستمعين الذين يتبقون بعد الفرز اللاحق إلى الحصول على الدرجات الشخصية المتوسطة.

وينبغي حساب الدرجة المتوسطة على النحو التالي: $$\hat{x} = \text{median}(x) = \begin{cases} x_{\frac{n+1}{2}} & n \text{ odd} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & n \text{ even} \end{cases}$$

وتُرتب x حسب الحجم.

وأول خطوة في التحليل هي حساب الدرجة المتوسطة $\overline{\eta}_{jk}$ لكل عرض من العروض. ويترتب على ذلك أن η_{ijk} هي الدرجة المتوسطة للملاحظ i في ظرف اختبار معين j وتتابع سمعي k وأن $\hat{\eta}$ هي متوسط العينة (جميع الملاحظين، وفي جميع الظروف، ولجميع التتابعات السمعية).

وبالمثل، يمكن حساب الدرجات المتوسطة الشاملة $\overline{\eta}_j$ و $\overline{\eta}_k$ لكل ظرف من ظروف الاختبار ولكل تتابع من تتابعات الاختبار.

وعلى الرغم من أن استعمال القيم المتوسطة ضروري في بعض طرائق التحليل مثل تحليل التباين (ANOVA) (انظر الفقرة 3.9)، فإن حساب المتوسط يعتبر إجراء بديلاً للنزعة المركزية. ويوفر المتوسط قياساً ثابتاً للنزعة المركزية يعتبر مثالياً للحالات التي تكون فيها مجموعة العينة صغيرة، أو يكون التوزيع فيها غير طبيعي، أو تحتوي فيها مجموعة البيانات على قيم شاذة ملحوظة. ويمكن أن تكون هناك عدة سيناريوهات للاختبار تكون فيها هذه الشواغل أقل تأثيراً. غير أنه، لأنّ من أعظم مزايا الاختبار الموحد مقارنة وتفسير الدرجات بين جميع المستعملين وأماكن الاستعمال، فمن المفيد تحديد طرائق للتحليل تكون الأكثر ثباتاً والأقل تأثراً بالعوامل التي قد تغير صلاحية الاختبار أو تقلل إمكانية الانتقال من اختبار إلى آخر.

وبهذه الطريقة يمكن تطبيق الإحصاءات غير المعْلَمية. وعند تطبيق تحليل البيانات غير المعْلَمي، ينبغي حساب المتوسطات وفواصل ثقة تبلغ نسبتها 95 في المائة باستعمال الطرائق المتاحة مثل استعمال خوارزمية إرجاع شائعة (bootstrapping).

ويمكن حساب قياسات الخطأ حول المتوسط باستعمال الانحراف المطلق للمتوسط:

$$\hat{\tau} = \Sigma |Y_i - \hat{\eta}| / n$$

ويوصى باستعمال المدى الرُبيعي (IQR) كقياس للثقة حول المتوسط. وهو الاختلاف في الدرجة بين الرُبيعين الأول والثالث: $IQR = Q_3 - Q_1$. وترد المعادلات في الفقرة 2.1.4. وإذا كان توزيع النتائج طبيعياً، فإن المدى الرُبيعي (IQR) يمثل ضعف الانحراف المطلق المتوسط.

ويوصَى بتحديد الدلالة الإحصائية عند مستوى دلالة 95 في المائة. كما أن الاختبارات غير المعْلَمية للعشوائية تعتبر قياسات ثابتة للدلالة الإحصائية. وخلافاً للتحليلات الإحصائية المعْلَمية، فإنها لا تستند إلى أي افتراضات بشأن التوزيع الأساسي للبيانات كما أنها أقل تأثراً بالكثير من الشواغل التي تتصل باستعمال حجم عينة أصغر.

ويسمح الاختبار غير المعْلَمي الثابت للعشوائية (اختبار التباديل) بتحديد احتمال حدوث اختلاف ملحوظ بين ظروف اختبارين إذا كانت البيانات عشوائية بحق كما يُفترض بموجب الفرضية الصفرية. ويعتبر الاحتمال المقيس في هذا الاختبار قياساً حقيقياً يُحدد من توزيع البيانات الفعلية وليس قياساً مستنتجاً يفترض شكلاً محدداً للتوزيع الأساسي [5]. ويتطلب هذا الشكل من الاختبار تقنيات شائعة لإعادة اختيار العينة مثل الإرجاع (bootstrapping) وتقنيات محاكاة مونت كارلو المتاحة حالياً بسهولة بسبب زيادة سرعة الحوسبة الحديثة [6]. ويرد في المرفق 3 وصف أثر تفصيلاً لطريقة الاختبار هذه.

2.9 تحليل القدرة

يمكن أن يكون تحليل القدرة مفيداً في تقدير أحجام العينة المطلوبة لاختبارات الاستماع إذا ما طبق كتحليل مسبق وفي تقدير القدرة أو الخطأ من النوع الثاني في الاختبار في تحليل لاحق. ويعطينا التحليل المسبق حجم العينة المطلوب للتجربة على أساس حجم تأثير $d = \frac{\bar{x}}{s}$، ومستوى دلالة α، وقدرة إحصائية $\beta - 1$.

وعلى العكس من ذلك يعطينا التحليل اللاحق قدرة $\beta - 1$ أو خطأ β من النوع الثاني في الاختبار على أساس حجم تأثير $d = \frac{\bar{x}}{s}$، ومستوى دلالة α، وحجم عينة N. وخطأ β من النوع الثاني هو احتمال وجود تأثير d في المجتمع الإحصائي ولكن لم يثبت

بالاختبار أنه ذو دلالة. وإذا زعم اختبار ما، مثلاً، أن الجودة لا تتأثر بالنظام، فإن β – 1 هو احتمال أن يكون الاختبار قد أثبت وجود انحطاط.[2]

3.9 تطبيق واستعمال تحليل التباين (ANOVA)

1.3.9 مقدمة

يركز هذا القسم على المتطلبات اللازمة لإجراء الإحصاءات المعْلَمية باستعمال تحليل التباين. ونظراً إلى ثبات نموذج تحليل التباين (انظر [7] و[8] و[12] و[13]) وقدرته الإحصائية[3]، فإنه يعتبر منهجية مناسبة جداً للبيانات التي يتم جمعها باستعمال منهجية التوصية ITU-R BS.1534. ونظراً إلى أن القيمة الإحصائية F في تحليل التباين تعتبر ثابتة إلى حد ما لكل من توزيعات البيانات غير الطبيعية وعدم تجانس التباين، فإن اختبار الافتراض يركز على طبيعة الخطأ أو القيم المتبقية.

ولقراءة المزيد عن الافتراضات العامة المرتبطة بالإحصاءات المعْلَمية، يرجى الرجوع إلى المرفق 4.

2.3.9 توصيف النموذج

يُنصح بشدة أثناء تصميم التجربة (انظر الفقرة 3)، بتوصيف النموذج بدقة فيما يتعلق بالمتغيرات المستقلة (مثلاً، العينة، والنظام، والظرف، وما إلى ذلك) والمتغيرات التابعة (مثلاً، الجودة السمعية الأساسية أو جهد الاستماع، وما إلى ذلك). كما ينبغي تعريف مستويات كل متغير مستقل في مرحلة توصيف النموذج.

وعند تعريف نموذج للتحليل (مثلاً، باستعمال تحليل التباين أو تحليل التباين بالقياسات المتكررة)، من المهم إدراج جميع المتغيرات المهمة. وقد يؤدى استبعاد المتغيرات المهمة، مثل تفاعلات العوامل المستقلة ذات الطريقتين أو الثلاث طرائق إلى سوء توصيف النموذج، مما يؤدي بدوره إلى ضعف تفسير التباين (R^2) واحتمال تفسير تحليل البيانات بشكل خاطئ.

3.3.9 قائمة التحليل الإحصائي المعْلَمي

تُعرض هذه القائمة كإرشاد موجز لاستعراض البيانات واختبار الافتراضات الأساسية (المعْلَمية وغير المعْلَمية) والخطوات الأساسية للإحصاءات المعْلَمية. وتركز القائمة على متطلبات تحليل التباين، كطريقة ملائمة لتحليل البيانات المستمدة من تجارب التوصية ITU-R BS.1534. وللاطلاع على دليل كامل يحال القارئ إلى كتب علم الإحصاء (مثلاً، [8] و[11] و[9]).

- الإحصاءات الاستطلاعية[4]
  - تأكد من أن هيكل البيانات صحيح وكما هو متوقع
  - ابحث عن البيانات الناقصة
  - ادرس طبيعة توزيع البيانات
  - راجع التوزيعات المحتملة الأخرى للبيانات (أحادية المنوال، وثنائية المنوال، والمتخالفة، وما إلى ذلك)
- أحادية البُعد
  - تأكد من استعمال المقيّمين المشترك للمقياس[5]
  - تأكد من أن البيانات ذات بُعد واحد في طبيعتها

[2] توجد أدوات كثيرة مثل G*Power [16] لإجراء تحليل القدرة بطريقة آلية لتوزيعات المجتمعات الإحصائية المعروفة في حين تكون أصعب لتوزيعات المجتمعات الإحصائية غير المعروفة.

[3] ينصح بشكل عام باختيار أقوى طريقة تحليل إحصائي تسمح بها البيانات [9] و[10].

[4] ينطبق هذا على الإحصاءات المعْلَمية وغير المعْلَمية بنفس القدر.

[5] لوحظ تعدد الأبعاد في الحالات التي تكون فيها لمجموعات فرعية من المجتمع الإحصائي آراء مختلفة بخصوص تقييم أصوات مصطنعة معينة.

– تحليل المكونات الرئيسية، مخططات تاكر–1 (Tucker-1) أو ألفا كرونباخ (Chronbach's alpha)

– استقلالية الملاحظات

– عادة ما يتم تعريف هذا في المنهجية التجريبية ولا يمكن اختبارها إحصائياً بسهولة. وينبغي التأكد من أن البيانات قد جمعت بشكل مستقل، أي باستعمال التقنيات التجريبية للحجب المزدوج والتأكد من أن المقيّمين لا يؤثرون على بعضهم البعض.

– تجانس التباين[6]

– اختبر افتراض أن كل متغير مستقل يُظهر تبايناً متشابهاً

– الاستعراض البصري باستعمال المخططات الصندوقية جنباً إلى جنب لكل مستوى من مستويات المتغيرات المستقلة؛ وكقاعدة بديهية، قد يختلف عدم التجانس كحد أقصى بمعامل قيمته 4

– يمكن استعمال اختبار براون وفورسايز (Brown and Forsythe) أو إحصاء ليفيني (Levene Statistic) لتقييم تجانس التباين

– التوزيع الطبيعي للقيم المتبقية

– اختبر التوزيع الطبيعي للقيم المتبقية

– اختبار دال لكولموغوروف–سميرنوف (Kolmogorov-Smirnov D test) أو اختبار ك–س لليلفورس (K-S Lillefors test) أو اختبار ليفيني (Levene's test)

– كما يمكن استعمال مخطط الاحتمال الطبيعي (يسمى أحياناً مخططات P-P) أو المخطط الخُمسي (يشار إليه كثيراً بمخططات Q-Q) كاختبار بصري للتوزيع الطبيعي

– اكتشاف القيم الشاذة

– ينبغي البحث عن القيم الشاذة ويمكن استبعادها عند وجود ما يبرر ذلك. وترد الإرشادات المتعلقة بهذه المسألة في الفقرة 2.1.4

– التحليل

– تحليل التباين – النموذج الخطي العام أو نموذج تحليل التباين بالقياسات المتكررة

– استعمل نموذج مناسب من نماذج ANOVA، مثلاً، النموذج الخطي العام (GLM) أو نموذج تحليل التباين بالقياسات المتكررة؛ ويرد مزيد من التفاصيل في المرفق 4

– قم بتوصيف النموذج وفقاً لتصميم التجربة

• أدرج تفاعلات ذات طريقتين وذات ثلاث طرائق حيث يكون ممكناً

– حلل البيانات مع النموذج والنتائج

• استعرض التباين المشروح (R^2) للنموذج المستعمل لوصف المتغير التابع

• راجع توزيع الخطأ المتبقي

• راجع العوامل الدالة وغير الدالة

– يمكن تكرار النموذج لإزالة القيم الشاذة والعوامل غير الدالة.

[6] مطلوب لتطبيق تحليل التباين ANOVA ولكن ليس لتحليل التباين بالقياسات المتكررة rmANOVA (انظر المرفق 4).

– الاختبارات اللاحقة

– طبق الاختبارات اللاحقة لإثبات دلالة الفرق بين المتوسطات حيث يكون العامل التابع (أو تفاعل العوامل) ذا دلالة في تحليل التباين.

– يتوفر عدد من الاختبارات اللاحقة المختلفة بمستويات مختلفة من التمييز، مثلاً، أقل فرق ذو دلالة لفيشر (LSD)، والفرق الدال بأمانة لتيوكي (HSD)، وما إلى ذلك.

– يُوصَى بالإبلاغ عن أحجام التأثير مع مستويات دلالتها.

– استخلاص النتائج

– بعد إجراء التحليل، قم بتلخيص النتائج بوضع مخططات للمتوسطات وفواصل ثقة تبلغ نسبتها 95 في المائة للبيانات الخام أو البيانات المعدة وفقاً لنموذج تحليل التباين (يشار إليها أحياناً بالمتوسطات الهامشية المقدرة).

– وفي الحالات التي يتبين فيها أن تفاعلات العوامل (ذات الطريقتين أو الثلاث طرائق) ذات دلالة، ينبغي رسم مخططات لها لإعطاء نظرة عامة شاملة على البيانات. وفي هذه الحالات سيوفر رسم مخططات للتأثيرات الرئيسية فقط نظرة عامة على البيانات يكون تأثير تفاعلها ملتبساً.

ويمكن الاطلاع على المزيد من الإرشاد حول استعمال نماذج تحليل التباين في المرفق 4 وفي النصوص الإحصائية والتطبيقية الشائعة، مثلاً، [11] و[13] و[15].

10 تقرير الاختبار وعرض النتائج

1.10 اعتبارات عامة

ينبغي أن يتم عرض النتائج بطريقة صديقة للمستعمل بحيث يستطيع القارئ سواء كان غير ذي خبرة أو كان خبيراً الحصول على المعلومات المهمة. وبدايةً، فإن أي قارئ يريد أن يرى النتيجة التجريبية العامة، ويفضل أن يكون ذلك في شكل بياني. ويمكن دعم هذا العرض بمعلومات كمية أكثر تفصيلاً، رغم أن التحليلات العددية التفصيلية الكاملة ينبغي أن ترد في مرفقات.

2.10 محتويات تقرير الاختبار

ينبغي أن ينقل تقرير الاختبار، كأوضح ما يكون، الأساس المنطقي للدراسة، والطرائق المستعملة، والنتائج التي تم التوصل إليها. ويجب تقديم تفاصيل كافية بحيث يستطيع الشخص العارف بالموضوع، من حيث المبدأ، أن يكرر الدراسة للتأكد تجريبياً من النتيجة. غير أنه من غير الضروري أن يحتوي التقرير على كل النتائج الفردية. وينبغي أن يكون القارئ المطلع قادراً على فهم التفاصيل الرئيسية للاختبار، مثل الأسباب الكامنة وراء الدراسة، وطرائق التصميم التجريبي والتنفيذ، والتحليلات والنتائج وإعداد تقرير ناقد عنها.

وينبغي إيلاء عناية خاصة للجوانب التالية:

– عرض بياني للنتائج؛

– عرض بياني لعملية فرز وتوصيف *المقيّمين ذوي الخبرة* الذين يتم اختيارهم؛

– تعريف التصميم التجريبي؛

– توصيف واختيار مادة الاختبار؛

– معلومات عامة عن النظام المستعمل لتجهيز مادة الاختبار؛

– تفاصيل تشكيل الاختبار؛

– التفسيرات المادية لبيئة الاستماع والأجهزة، بما في ذلك أبعاد الغرف وخصائصها السمعية، وأنواع وأماكن محولات القدرة، ومواصفات الأجهزة الكهربائية (انظر الملاحظة 1)؛

- التصميم التجريبي، والتدريب، والتعليمات، والتتابعات التجريبية، وإجراءات الاختبار، وتوليد البيانات؛
- معالجة البيانات، بما في ذلك تفاصيل الإحصاءات الاستنتاجية التحليلية والوصفية؛
- استعمال المرتكزات في الاختبار؛
- طرائق الفرز اللاحق المستعملة في تحليل النتائج – ويشمل هذا طرائق استبعاد القيم الشاذة أو المستمعين غير المدربين؛
- تحديد ما إذا كان الاختبار قد نفذ باستعمال التوصية ITU-R BS.1534 أو التوصية ITU-R BS.1534-1؛ وينبغي ذكر ذلك بوضوح في الوثيقة مع وصف ظروف المرتكز المستعمل؛
- التعريف المناسب ورمز التوليد اللازم للسماح للمستعمل الجديد بإنتاج أي مرتكز مستعمل في الاختبار ولا يرد وصفه صراحة في التوصية ITU-R BS.1534-2؛
- الأساس التفصيلي لكل النتائج المستخلصة.

الملاحظة 1– نظراً إلى أن هناك بعض الأدلة على أن ظروف الاستماع، مثلاً، النسخ بمكبر الصوت مقابل النسخ بسماعة الرأس، قد توثر على نتائج التقييمات الشخصية، يُطلب من القائمين بالتجربة الإبلاغ صراحةً عن ظروف الاستماع، ونوع أجهزة النسخ المستعملة في التجارب. وإذا كان المقصود إجراء تحليل إحصائي مركب لأنواع محولات القدرة المختلفة، ينبغي التأكد مما إذا كان تركيب النتائج على هذا النحو ممكناً.

3.10 عرض النتائج

يجب ذكر المتوسط والمدى IQR للتوزيع الإحصائي لدرجات التقييم لكل مُعْلَمة من مُعْلَمات الاختبار .

ويجب أن تقدم النتائج مع المعلومات التالية:

- وصف مواد الاختبار؛
- عدد المقيّمين؛
- يجب إدراج عرض بياني للنتائج؛ ومخططات صندوقية توضح مدى IQR، بالإضافة إلى عرض المتوسطات وفواصل ثقة تبلغ نسبتها 95 في المائة؛ وينبغي الإبلاغ عن الاختلافات الدالة بين الأنظمة التي يجري اختبارها وطريقة التحليل الإحصائي المطبقة.

وبالإضافة إلى ذلك، يمكن أيضاً عرض النتائج في أشكال ملائمة مثل المتوسطات وفواصل الثقة عندما تدعم البيانات هذه العروض بعد العرض التصويري لمخططات الصناديق.

4.10 الدرجات المطلقة

يعطي عرض الدرجات المتوسطة المطلقة للأنظمة الخاضعة للاختبار، والمرجع المحجوب، والمرتكزات نظرة عامة جيدة على النتيجة. غير أنه يتعين على المرء أن يتذكر دائماً أن ذلك لا يعطيه أي معلومات عن التحليل الإحصائي المفصّل. وبالتالي، فإن الملاحظات تكون غير مستقلة ولن يسفر التحليل الإحصائي للدرجات المطلقة فقط، بدون اعتبار للمجتمع الإحصائي الأساسي للعينة التي تجري ملاحظتها، عن أي معلومات ذات مغزى. وبالإضافة إلى ذلك، ينبغي الإبلاغ عن الطرائق الإحصائية المطبقة على النحو المقترح في الفقرة 9.

5.10 مستوى الدلالة وفاصل الثقة

ينبغي أن يزود تقرير الاختبار القارئ بمعلومات عن الطبيعة الإحصائية الأصيلة لجميع البيانات الشخصية. وينبغي ذكر مستويات الدلالة، وكذلك التفاصيل الأخرى المتعلقة بالطرائق الإحصائية والنواتج التي ستسهل الفهم على القارئ. ويمكن أن تشمل هذه التفاصيل فواصل الثقة أو أعمدة الخطأ في أشكال بيانية.

وبالطبع، فليس هناك مستوى دلالة "صحيح"، ويتم عادةً اختيار القيمة 0,05. ويمكن، من حيث المبدأ استعمال اختبار أحادي الذيل أو ثنائي الذيل تبعاً للفرضية التي يجري اختبارها.

المراجع

[1] Stevens, S. S. (1951). Mathematics, measurement and psychophysics, in Stevens, S. S. (ed.), Handbook of experimental psychology, John Wiley & Sons, New York.

[2] EBU [2000a] MUSHRA – Method for Subjective Listening Tests of Intermediate Audio Quality. Draft EBU Recommendation, B/AIM 022 (Rev.8)/BMC 607rev, January.

[3] EBU [2000b] EBU Report on the subjective listening tests of some commercial internet audio codecs. Document BPN 029, June.

[4] Soulodre, G. A., & Lavoie, M. C. (1999, August). Subjective evaluation of large and small impairments in audio codecs. In *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*. Audio Engineering Society.

[5] Berry, K. J., Johnston, J. E., & Mielke, P. W. (2011). Permutation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(6), 527-542.

[6] Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, 38.

[7] Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. Lawrence Erlbaum Associates, Inc.

[8] Keppel, G. and Wicken., T. D. (2004). Design and Analysis. *A Researcher's Handbook*, 4th edition. Pearson Prentice Hall.

[9] Garson, D. G. Testing statistical assumptions, Blue Book Series, Statistical Associates Publishing, 2012.

[10] Ellis, P. D. (2010). The essential guide to effect sizes. *Cambridge: Cambridge University Press*, *2010*, 3-173.

[11] Howell., D.C. (1997). Statistical methods for psychology, 4th Edition, Duxbury Press.

[12] Kirk., R.E., (1982). Experimental Design: Procedures for the Behavioural Sciences, 2nd edition. Brooks/Cole Publishing Company 1982.

[13] Bech, S., & Zacharov, N. (2007). Perceptual audio evaluation-Theory, method and application. John Wiley & Sons.

[14] Khan, A. and Rayner, G. D. (2003). Robustness to Non-Normality of Common Tests for the Many-Sample Location Problem, *Journal of Applied Mathematics & Decision Sciences*, 7(4), 187-206.

[15] ITU-T. Practical procedures for subjective testing, International Telecommunication Union, 2011.

[16] Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*,*(4)*, 1149-1160.

المرفق 1
بالملحق 1
(معياري)

التعليمات التي يتعين إعطاؤها للمُقيّمين

التعليمات التالية مثال على نوع التعليمات التي ينبغي إعطاؤها للمقيّمين أو قراءتها عليهم لتعريفهم بطريقة إجراء الاختبار.

1 مرحلة التعرّف أو التدريب

الخطوة الأولى في اختبارات الاستماع هي أن يصبح المرء معتاداً على عملية الاختبار. وتُسمى هذه المرحلة مرحلة التدريب وتسبق مرحلة التقييم الرسمي.

والغرض من مرحلة التدريب هو أن تتيح لك، بوصفك مُقيّماً، تحقيق الهدفين التاليين:

- **الجزء ألف:** التعرّف على جميع المقتطفات السمعية الخاضعة للاختبار ومديات سويات جودتها؛
- **الجزء باء:** تعلُّم استعمال أجهزة الاختبار ومقياس تحديد الدرجات.

وفي الجزء ألف من مرحلة التدريب، ستستطيع الاستماع إلى جميع المقتطفات الصوتية التي تم اختيارها للاختبارات لتمثيل المدى الكامل لدرجات الجودة المحتملة. وسوف تكون الفقرات الصوتية، التي ستستمع إليها، حرجة إلى حد ما تبعاً لمعدل البتات و"الظروف" الأخرى المستعملة. ويوضّح الشكل 3 السطح البيني للمستعمل. ويمكنك النقر على المفاتيح المختلفة للاستماع إلى مقتطفات سمعية مختلفة بما فيها المقتطفات المرجعية. وبهذه الطريقة يمكنك أن تتعلم كيف تتذوّق مدى مستويات جودة مختلفة لفقرات البرامج المختلفة. ويتم تجميع المقتطفات على أساس الظروف المشتركة. ويتم في هذه الحالة تحديد ثلاث من هذه المجموعات. وتشمل كل مجموعة أربع إشارات مُعالَجة.

وفي الجزء باء من مرحلة التدريب، ستتعلم استعمال أجهزة إعادة التشغيل وتسجيل الدرجات المتاحة التي ستستعمل في تقييم جودة المقتطفات الصوتية.

وينبغي أن تتعلم أثناء مرحلة التدريب كيف تفسر، كفرد، مظاهر الانحطاط المسموعة في حدود مقياس تحديد الدرجات. ولا ينبغي أن تناقش تفسيرك الشخصي للمقياس مع المقيّمين الآخرين في أي وقت أثناء مرحلة التدريب. غير أنه من المستحب أن تشرح الأصوات المصطنعة للمقيّمين الآخرين.

ولن تؤخذ في الحسبان في الاختبارات الحقيقية أي درجات تُعطَى أثناء مرحلة التدريب.

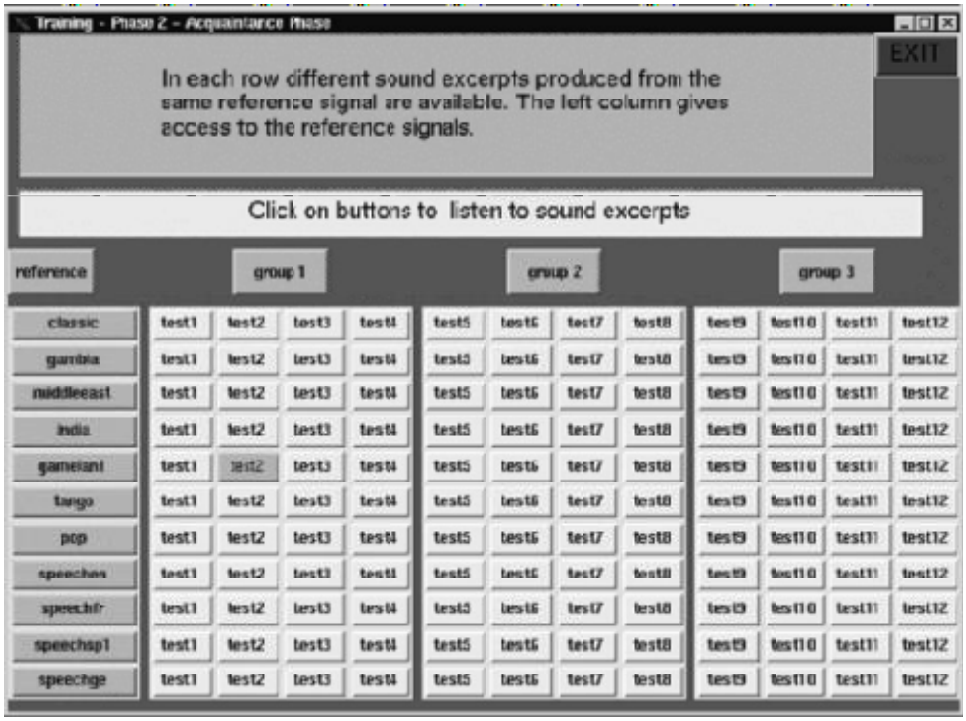
2 مرحلة تحديد درجات المواد وهي محجوبة

الغرض من تحديد الدرجات والمواد محجوبة دعوتك إلى إعطاء درجات باستعمال مقياس الجودة. وينبغي أن تعبّر درجاتك عن حُكمك الشخصي على سوية جودة كل مقتطف من المقتطفات السمعية المعروضة عليك. وسوف تحتوي كل محاولة على 9 إشارات يجب تحديد درجاتها. ويبلغ طول كل فقرة 10 s تقريباً. وينبغي أن تستمع إلى المرجع، والمرتكز، وكل ظروف الاختبار بالنقر على المفاتيح المخصصة لذلك. ويمكنك الاستماع إلى الإشارات بأي ترتيب، وفي أي عدد من المرات.

واستعمل المسطرة الزلاقة لتتوقف عند أي إشارة لتوضح رأيك في جودتها. وعندما تشعر بالرضى عن الدرجات التي حددتها لكل الإشارات يتعين عليك أن تقوم بالنقر على مفتاح "سَجّل الدرجات" (register scores) أسفل الشاشة.

الشكل 3

صورة توضح مثالاً على السطح البيني للمستعمل للجزء ألف من مرحلة التدريب



وعليك أن تستعمل مقياس الجودة على النحو الوارد في الشكل 1 عند إعطاء درجاتك.

ويستمر مقياس تحديد الدرجات من "ممتازة" إلى "رديئة". وتوافق الدرجة 0 أدنى درجات الفئة "رديء"، في حين توافق الدرجة 100 أعلى درجات الفئة "ممتاز".

وعند تقييمك للمقتطفات الصوتية، يُرجَى ملاحظة أنه لا ينبغي لك بالضرورة أن تُعطِي درجة في الفئة "رديء" للمقتطف الصوتي ذي الجودة الدنيا في الاختبار. غير أنه يجب إعطاء الدرجة 100 لمقتطف واحد أو أكثر نظراً إلى إدراج الإشارة المرجعية غير المعالَجة بوصفها واحداً من المقتطفات التي يجب تحديد درجاتها.

الشكل 4

على السطح البيني للمستعمل في مرحلة تحديد الدرجات والأصوات محجوبة

المرفق 2
بالملحق 1
(إعلامي)

ملاحظات توجيهية بشأن تصميم السطح البيني للمستعمل

المقترحات التالية موجّهة لأولئك الذين قد يفكرون في:

أ) إنتاج أنظمة لإجراء الاختبارات الشخصية وفقاً لطريقة MUSHRA؛

ب) إجراء هذه الاختبارات.

والهدف من هذه الاقتراحات هو زيادة موثوقية نتائج الاختبارات وتيسير تحليل أي حالات شاذة قد تظهر أثناء معالجة درجات الاختبار.

وينبغي أن يصمم السطح البيني للمستعمل بطريقة تقلل من احتمال أن يعطي الفرد المشارك في الاختبار درجة لا تعبر عن قصده الحقيقي. ولهذا الغرض، ينبغي أن تتخذ الخطوات اللازمة لكفالة أن يكون واضحاً من السطح البيني للمستعمل أي نُسخ من مواد الاختبار المعالَجة هي التي يستمع إليها الفرد في وقت معين. ويمكن مساعدة هذه العملية عن طريق اختيار الألوان ونصوع المؤشرات على الشاشة بعناية (مثل وجود مفاتيح يمكن النقر عليها) لتجنب الصعوبات المحتملة التي قد تنشأ في حالة عدم تأثر الفرد ببعض الألوان.

وينبغي أيضاً كفالة أن يكون الفرد قادراً على تحديد الدرجة المعطاة للمادة التي يتم الاستماع إليها في الوقت الحاضر فقط. وقد لوحظ أن بعض المقيّمين يستمعون إلى نسختين معالجتين للمادة الواحدة، على التوالي، لإعطاء درجة للأولى وليس للأخيرة، التي يسمعونها. وفي هذه الحالة، يحتمل ارتكاب خطأ (وخاصة عند عرض عدد كبير من أدوات التحكم على الشاشة) فقد تعطى الدرجة لإشارة أخرى غير الإشارة المقصودة. ولمحاولة خفض هذا الاحتمال، يُقترح أن تكون وسيلة التحكم الوحيدة المَنَشّطة في أي وقت من الأوقات هي الوسيلة المتعلقة بالإشارة التي يتم الاستماع إليها في الوقت الحاضر. وينبغي إيقاف عمل وسائل التحكم في إعطاء الدرجات للإشارات الأخرى، التي لا يتم الاستماع إليها في الوقت الحاضر.

المرفق 3
بالملحق 1
(معياري)

وصف للمقارنة الإحصائية غير المعْلَمية بين عينتين باستعمال تقنيات إعادة اختيار العينة وطرائق محاكاة مونت كارلو

يمكن استعمال الاختبارات غير المعْلَمية للعشوائية مع التقنيات الشائعة لإعادة اختيار العينة مثل إجراءات الإرجاع (bootstrapping) لتقرير دلالة أي نتيجة إحصائية تقريباً. وعلى سبيل المثال، يمكن حساب دلالة الفرق المتوسط في الاستجابة الملحوظة بين إشارتي اختبار (بأحجام العينة =*N1* و*N2*) بالطريقة التالية: يدون الفرق الفعلي بين متوسطي كل عينة ويشار إليه بوصفه $Diff_{ACT\_1}$. ثم يتم بعد ذلك تجميع كل البيانات المستمدة من هذه العينات في ملف واحد أو مُتّجَه. ويستعمل إجراء الإرجاع بحيث تتم تباديل التراكم في كل تكرار بالعينات المستمدة من الحجم *N1* و*N2* بدون إحلال. ويُسجل الفرق بين متوسطي العينتين المختارتين بطريقة عشوائية بوصفه $Diff_{EST\_1}$. ويمكن بعد ذلك تكرار هذا الإجراء 000 10 مرة وسوف ينتُج عن نسبة عدد المرات التي زاد فيها $Diff_{EST\_N}$ على $Diff_{ACT\_N}$ مقسوماً على 000 10 قيمة *P* المقابلة. وإذا كان العدد الكلي للمرات التي يزيد فيها $Diff_{EST\_N}$ عن $Diff_{ACT\_N}$ أقل من 500 (0,05 = 500/10 000)، يمكن أن يقال إن الفرق بين المتوسطين هو فرق دال عند مستوى 0,05، $p < 0,05$.

المرفق 4
بالملحق 1
(إعلامي)

ملاحظات توجيهية بشأن التحليل الإحصائي المعْلَمي

1 مقدمة

يرد في الفقرة 9 وصف للتحليل الإحصائي المعْلَمي الأساسي للنتائج في اختبارات MUSHRA. غير أنه، وبخاصة عندما يكون المطلوب مقارنة ظروف كثيرة كل منها بالآخر، يفضل اختبار شامل مثل تحليل التباين (ANOVA) على المقارنات الزوجية المتعددة. ويصف هذا المرفق كيف يمكن القيام بذلك. ويشمل الشروط المسبقة للتحليل ويبرز البدائل عند عدم استيفاء هذه الشروط.

ويَستعمل اختبار MUSHRA *مقاييس متكررة* أو *تصميماً داخل مشاهدات العينة* (يمكن الاطلاع على مقدمة ممتازة لهذين المفهومين في Maxwell & Delaney, 2004) حيث يوجد تداخل كامل بين اثنين من العوامل الموجودة داخل مشاهدات العينة (الظرف والمادة السمعية)، ويتم الحصول على تقييم واحد على الأقل لكل مجموعة من المستمعين والمادة السمعية والظرف. كما يمكن أن توجد حالات تعرض فيها نفس التركيبات المكونة من المادة السمعية والظرف على اثنتين أو أكثر من مجموعات المقيّمين المختلفة، مثلاً في مختبرين مختلفين. وفي هذه الحالة يوجد عامل إضافي بين مشاهدات العينة يرجع إلى *المجموعة* وينبغي أن يُراعى في التحليل.

وتعتبر الإحصاءات الاستنتاجية ضرورية لتعميم النتائج التي يتم الحصول عليها في عينة صغيرة نسبياً في المستمعين على مجتمع كل المستمعين. وعلى سبيل المثال، إذا كانت التقييمات في اختبار الاستماع تشير إلى وجود فرق بين الجودة السمعية المدركة لأي مشفر جديد والمشفر القائم، فمن المهم الإجابة على السؤال المتعلق بما إذا كان يمكن توقع هذا الفرق إذا قيّمت مجموعة مختلفة تماماً من المستمعين الجودة السمعية للنظاميين. وفيما يتعلق بالتصميم المحدد لاختبارات استماع MUSHRA، هناك ثلاثة أسئلة على الأقل قد يرغب المرء في الإجابة عليها (أو بالمصطلحات الإحصائية، فرضيات يرغب المرء في اختبارها)، وتقدم الإحصاءات الاستنتاجية الوارد وصفها هنا إجابات صالحة. أولاً، سيكون السؤال المتعلق بالاهتمام الرئيسي هو ما إذا كانت الجودة السمعية المدركة قد اختلفت بين الأنظمة الخاضعة للاختبار (مثلاً، مرجع وثلاثة مشفرات مختلفة). وثانياً، إذا تم تقييم الأنظمة السمعية في اختبار الاستماع باستعمال مواد اختبار مختلفة، هل اعتمدت عمليات تقييم الجودة السمعية على المادة السمعية؟ وثالثاً، هل اختلف تأثير النظام السمعي على الجودة السمعية المدركة بين مواد الاختبار؟ والطريقة الملائمة للإجابة على هذه الأسئلة تتمثل أولاً في الحصول على اختبارات دلالة للتأثير الرئيسي للظرف (النظام السمعي) والتأثير الرئيسي للمادة السمعية وظرف تفاعل المادة السمعية x بإجراء تحليل التباين. ويكون هناك تفاعل عندما تعتمد الاختلافات بين الجودة المدركة للأنظمة السمعية على المادة السمعية. ويلاحظ أنه بسبب التفاعلات المحتملة لا ينصح بتجميع التقييمات الخاصة بكل نظام سمعي في جميع المواد السمعية، حتى إذا لم يكن المرء مهتماً بشكل خاص بتأثير المادة السمعية أو تأثير التفاعل. ثم يمكن اختبار فرضيات أكثر تحديداً، تتعلق مثلاً بالفرق المدرك بين زوج من الأنظمة السمعية، وذلك باستعمال مقارنات إضافية.

وحيثما يجب مقارنة أكثر من ظرفين تجريبيين، كأربعة مشفرات مختلفة مثلاً، فليس من الملائم تأسيس الإحصاءات الاستنتاجية على المقارنات الزوجية المتعددة. وعلى سبيل المثال، إذا أُدرِجت خمسة أنظمة سمعية في الاختبار، أي $K = 5$ (4 مشفرات بالإضافة إلى المرجع)، فإن هناك $\binom{K}{2} = K(K-1)/2 = 10$ أزواج من الظروف. وسوف يؤدي اختبار الفوارق في كل من هذه الأزواج العشرة باستعمال اختبارات t القائمة على العينات الزوجية العشرة عند مستوى α يساوي 0,05 إلى تضخم ما يطلق عليه اسم معدل الخطأ من النوع الأول. وفي كل اختبار t فردي، يكون احتمال الرفض الخاطئ للفرضية الصفرية بعدم وجود فوارق بين الجودة السمعية المدركة في المشفرين هو α.

وفي عدد C من هذه الاختبارات، فإن احتمال ارتكاب خطأ واحد على الأقل من النوع الأول قدره $1 - (1 - \alpha)^C$، وهو في مثالنا حيث $C = 10$ يساوي 0,40 وهو بذلك أعلى كثيراً من مستوى α المطلوب وهو 0,05. ويمكن التحكم في معدل الخطأ المجموعة بتطبيق التصحيحات الملائمة للاختبار المتعدد مثل تصحيح بونفيروني (Bonferroni) أو إجراء هوشبرغ (Hochberg) لعام 1988 الوارد وصفه في موضع لاحق. غير أن اختبارات t الزوجية حتى مع التصحيح تخفي المعلومات المهمة، في جزء منه لأن اختبارات t المتعددة على جميع أزواج المتوسطات تستعمل معلومات متكررة (كل متوسط يظهر في عدة اختبارات). وعادة ما سيكون نهج الاختبار الزوجي أقل قدرة (أي أقل حساسية في اكتشاف الفرق بين الظروف) من استعمال الاختبار الشامل الملائم، الذي يكون في حالة اختبار MUSHRA تحليل التباين بالقياسات المتكررة. ويرد فيما يلي وصف خطوة بخطوة لتحليل البيانات لحالة يطبق فيها اختبار MUSHRA لا تحتوي على عوامل بين مشاهدات العينة. وبعبارة اخرى، يُفترض اختبار مجموعة واحدة فقط من المقيّمين وأن كل تركيبات الظرف والمادة السمعية عرضت على كل مقيّم مرة واحدة على الأقل. وسوف يرد في موضع لاحق وصف لتوسيع التصميم باستعمال أكثر من مجموعة واحدة (مثلاً عندما يجرى الاختبار في مختبرين).

2 اختبار التوزيع الطبيعي

من الحكمة أن توضع في الاعتبار تأثيرات الانحراف المحتمل لقياس الأجوبة عن التوزيع الطبيعي على صحة الاختبار الإحصائي. وبالنسبة لتصميم بين مشاهدات العينة يتم فيه اختبار كل مقيّم في ظرف تجريبي واحد، تعتبر تحليلات التباين المنفذة في إطار النموذج الخطي العام ثابتة بدرجة مدهشة في ضوء عدم التوزيع الطبيعي لقياس الأجوبة (مثلاً، [11] و[13] و[25] و[35]).

وبالنسبة لتصميم القياسات المتكررة مثل اختبار MUSHRA، نلاحظ أولاً طريقة بديلة لاختبار الفرضية الصفرية التي تفيد بأن الجودة السمعية المدركة متماثلة في جميع الظروف في المجتمع الإحصائي. ويعتبر هذا معادلاً لحساب تناقضات عمودية عددها $K\text{-}1$، مثلاً عن طريق تكوين متغيرات الفرق بين الظروف البالغ عددها K، ثم اختبار فرضية أن متوسط المجتمع الإحصائي لكل متغيرات الفروق هذه يساوي 0. فإذا كان الاختبار يشمل، مثلاً، المرجع ومُشفّرين، يمكن إنشاء اثنين من متغيرات الفرق D_1 وD_2 بحساب الفرق بين تقييم المرجع وتقييم المشفّر A (D_1) والفرق بين تقييم المشفّر A وتقييم المشفّر B (D_2). وتفترض جميع نُهُج تحليل التباين بالقياسات المتكررة أن كل متغيرات الفرق هذه موزعة توزيعاً طبيعياً. ومن المؤسف أن انعدام التوزيع الطبيعي، خلافاً للتصميم بين مشاهدات العينة، يمكن أن ينتج عنه معدلات خطأ من النوع الأول متحفظة للغاية أو متحررة للغاية ([5]؛ و[22]؛ و[30]؛ و[39]). وهذا يعني أنه بالنسبة لمستوى α معين (مثلاً، $\alpha = 0{,}05$) ستكون نسبة الحالات التي يُنتِج فيها تحليل التباين قيمة p دالة إحصائياً ($p < \alpha$) على الرغم من أن الفرضية الصفرية التي تفيد بتماثل المتوسطات في جميع الظروف صحيحة، أقل أو أعلى من القيمة الاسمية α. ومرة أخرى، خلافاً للتصميم بين مشاهدات العينة، فإن مجرد زيادة حجم العينة لا يحل هذه المشكلة [30]. وثمة أدلة متزايدة على أن الابتعاد عن التناظر له تأثير أكثر خطورة من حالات الانحرافات عن التوزيع الطبيعي فيما يتعلق بالتفرطح ([4]؛ و[18]). ويمكن التعبير عن درجة الانحراف عن التناظر من حيث *تخالف* التوزيع وهو العزم المنمط الثالث [8]. وفيما يتعلق بتوزيع تناظري مثل التوزيع الطبيعي، يكون التخالف قدره 0. والتفرطح هو العزم المنمط الرابع للمجتمع الإحصائي حول المتوسط ويصف أوزان الذروة والذيل (انظر [9] للاطلاع على أمثلة). وتشير دراسات المحاكاة السابقة إلى أنه فيما يتعلق بالانحرافات الصغيرة عن التناظر، سوف تضبط تحليلات التباين بالقياسات المتكررة (rmANOVA) معدل الخطأ من النوع الأول. غير أن الحالة الراهنة للبحوث لا تسمح بصياغة قواعد دقيقة تتعلق بالدرجة المقبولة للانحراف عن التوزيع الطبيعي. ولذلك، يوصى باختبار وجود التوزيع الطبيعي متعدد المتغيرات، والإبلاغ عن التقديرات التجريبية للانحراف والتفرطح.

ومن المهم ملاحظة أن النموذج الخطي العام الكامن وراء تحليلات التباين بالقياسات المتكررة لا يفترض أن الأجوبة الخام (أي التقييم في اختبار MUSHRA) موزعة توزيعاً طبيعياً. وبدلاً من ذلك، فإن النموذج يفترض أن *الأخطاء* موزعة توزيعاً طبيعياً. ولهذا السبب، يجب حساب اختبارات التوزيع الطبيعي أو قياسات الانحراف والتفرطح *للقيم المتبقية* في النموذج، وليس للبيانات الخام. ومن حسن الحظ أن معظم البرمجيات الإحصائية قادرة على حفظ القيم المتبقية لكل ظرف تجريبي يتم تحليله، وهو في هذه الحالة كل تركيب من الأنظمة السمعية والمادة السمعية. وسوف يوفر هذا متجهاً للقيم المتبقية لكل ظرف تجريبي. وفي كل متجه، تمثل كل قيمة مقيّماً واحداً.

وتتوفر اختبارات عديدة للتوزيع الطبيعي متعدد المتغيرات مثل اختبار شابيرو–ويلك (Shapiro-Wilk) متعدد المتغيرات الذي اقترحه رويستون (Royston) [34] واختبارات قائمة على التخالف والتفرطح متعددة المتغيرات [10] ونُهُج أخرى [14]. وتتوفر تعليمات تشغيل (macros) لتطبيق هذه الاختبارات في برنامج SPSS (http://www.columbia.edu/~ld208/normtest.sps) وبرنامج SAS (http://support.sas.com/kb/24/983.html). ومن المحتمل إلى حد كبير أيضاً في حزم برمجيات أخرى. وتتوفر في جميع حزم البرمجيات الإحصائية الرئيسية تقديرات أحادية المتغيرات للتخالف والتفرطح يمكن حسابها بشكل مستقل بالنسبة للقيم المتبقية في كل تركيبة للأنظمة السمعية والمادة السمعية. كما تَحسب تعليمات التشغيل التي أعدها DeCarlo [9] (http://www.columbia.edu/~ld208/normtest.sps) أيضاً التخالف والتفرطح متعددي المتغيرات [26]. ويجب الإبلاغ عن تقديرات التخالف والتفرطح أحادية المتغيرات أو متعددة المتغيرات، وكذلك عن نتيجة اختبار التوزيع الطبيعي متعدد المتغيرات.

وإذا لم يكن اختبار التوزيع الطبيعي متعدد المتغيرات دالاً، أو إذا لم تُظهر الاختبارات متعددة المتغيرات أو أحادية المتغيرات أي انحراف دال في التخالف أو التفرطح عن القيم المتوقعة للتوزيع الطبيعي، فإن افتراضات تحليل التباين بالقياسات المتكررة تكون قد تحققت.

غير أنه إذا أشارت أي اختبارات إلى وجود انحراف دال عن التوزيع الطبيعي، أو إذا تجاوز التخالف في أي ظرف تجريبي قيمة 0,5 (كقاعدة بديهية أولية)، فإن ذلك يثير سؤالاً عما ستكون عليه التبعات المترتبة على هذه النتائج. وهناك مشكلتان عامتان، يرتبط كلاهما بما جرت مناقشته عن عدم وجود قواعد تتعلق بالانحراف المقبول عن التوزيع الطبيعي بالنسبة إلى تحليلات التباين بالقياسات المتكررة. والمشكلة الأولى هي أن اختبارات التوزيع الطبيعي متعدد المتغيرات حساسة إلى حد ما، وسوف تكتشف في الغالب الانحرافات الصغيرة للغاية عن التوزيع الطبيعي. كما أنها لن تكتشف انعدام التناظر في توزيع القيم المتبقية فحسب، بل أيضاً التفرطح أو أي جوانب أخرى في التوزيع سوف يؤدي دوراً، في حين أنه من المحتمل إلى حد كبير أن ينتج عن انعدام التناظر معدلات غير ثابتة للخطأ من النوع الأول في تحليلات التباين بالقياسات المتكررة. وثانياً، إذا تم تقدير قياسات التخالف والتفرطح متعددي المتغيرات من البيانات [26]، فإن هذه المعلومات لا تسمح باتخاذ قرار حول ما إذا كان يمكن تطبيق تحليل التباين بالقياسات المتكررة، مرة اخرى بسبب عدم وجود قواعد تتعلق بالانحراف المقبول عن التوزيع الطبيعي. ويؤكد هذا الحاجة إلى الإبلاغ عن قياسات التخالف والتفرطح وكذلك نتائج الاختبار. وبمجرد توافر قواعد صالحة تتعلق بالانحراف المقبول عن التوزيع الطبيعي، يمكن عندئذ إعادة تقييم نتائج اختبارات تحليل التباين بالقياسات المتكررة باستعمال المعلومات المحسنة. وإذا بدا أن الانحراف عن التوزيع الطبيعي شديد، كأن يشار إليه مثلاً بتقديرات للتخالف تزيد عن 1,0 [29]، يمكن في هذه الحالة النظر في استعمال البدائل غير المعْلَمية لتحليل التباين بالقياسات المتكررة، مثل الاختبارات التي تستعمل تقنيات إعادة اختيار العينة أو اختبار فريدمان (Friedman). غير أنه لم يتضح حتى الآن الحالات التي ستحل فيها تقنيات إعادة اختيار العينة مشكلة انعدام التوزيع الطبيعي [38]. ولا يفترض اختبار فريدمان التوزيع الطبيعي متعدد المتغيرات، ولكنه يفترض أن التباينات متشابهة في كل الظروف التجريبية [36]، وفي الغالب لن يكون ذلك هو الحال بالنسبة للبيانات التجريبية. وعلاوة على ذلك، فإن اختبار فريدمان هو اختبار أحادي المتغيرات. ولذلك، فحتى إذا تم تحقيق افتراض تساوي التباينات، يمكن استعمال اختبار فريدمان لاكتشاف تأثير للنظام السمعي موزع في المتوسط عبر المادة السمعية، ولكن لا يمكن استعماله في تحليل تفاعل المادة السمعية في النظام السمعي.

3 اختيار نهج تحليل التباين بالقياسات المتكررة

بالنسبة للبيانات المستمدة من تصميم القياسات المتكررة، توجد نُهُج مختلفة كثيرة لاختبار تأثيرات العوامل الموجودة داخل مشاهدات العينة وفيما بينها [21]. ونظراً إلى أننا ننظر الآن في حالة تصميم لا يحتوي على عوامل (تجميع) بين مشاهدات العينة، ونظراً إلى أننا نفترض عدم وجود بيانات ناقصة (أي أن هناك تقييماً لكل تركيبة مكونة من المستمع والمادة السمعية والظرف)، فإن هناك نهجين يمكن أن يوصَى بهما. ويوفر كلاهما اختبارات صالحة للفرضيات عندما تتسم البيانات بالتوزيع الطبيعي متعددة المتغيرات ولكنهما قد يختلفا في قدرتهما الإحصائية (أي الحساسية في اكتشاف الابتعاد عن الفرضية الصفرية)، وفقاً لعوامل أخرى من بينها حجم العينة.

وأسلوبا التحليل هما (أ) *النهج أحادي المتغيرات مع تصحيح هيونه–فلدت لدرجات الحرية*، و(ب) *النهج متعدد المتغيرات*. ويمكن الاطلاع على أوصاف هذين النهجين في مواضع أخرى [21]؛ و[28]. والأسلوبان متوفران في الحزم البرمجية الإحصائية الرئيسية (مثل R، وSAS، وSPSS، وStatistica).

وبسبب هيكل البيانات القائم على القياسات المتكررة، فإن هناك ارتباطاً بين التقييمات التي يتم الحصول عليها في التركيبات المختلفة للظرف والمادة السمعية. وعلى سبيل المثال، إذا أعطى المستمع تقييماً مرتفعاً بشكل غير عادي للمرتكز منخفض الجودة، فإن تقييماته للمشفرات ستميل أيضاً إلى أن تكون أعلى من تقييمات المقيّمين الآخرين. ويفترض نهج المتغير الأحادي أن يكون هيكل التباين–التباين المشترك للبيانات دائرياً، وهو ما يعادل القول إن متغيرات الفروق الموصوفة أدناه لها كلها نفس التباين [16]؛ و[33]. غير أن هذا الافتراض لا يتحقق فعلياً لكل مجموعات البيانات التجريبية تقريباً [21]. ولحل هذه المشكلة، يطبق عامل تصحيح على درجات الحرية عند حساب قيمة p وفقاً للتوزيع F. وتحقيقاً لهذا الهدف، يتم تقدير كمية الابتعاد عن الدائرية من البيانات. ويوصَى باستعمال عامل تصحيح هيونه–فلدت (Huynh-Feldt) الذي يرمز إليه بالمصطلح [17] $\tilde{\varepsilon}$ لأن عامل تصحيح غرينهاوس-غايسر البديل [12] يميل إلى إنتاج اختبارات متحفظة (مثلاً، [17]؛ و[30]). وعندما يكون توزيع البيانات توزيعاً طبيعياً يُنتج النهج أحادي المتغيرات باستعمال تصحيح هيونه–فلدت معدلات صالحة للخطأ من النوع الأول حتى بالنسبة لأحجام العينة شديدة الصغر ($N = 3$). وتوفر كل حزم البرمجيات الإحصائية الرئيسية عامل التصحيح $\tilde{\varepsilon}$ وقيم p المصححة.

ويستعمل *النهج متعدد المتغيرات* صيغة بديلة ولكنها مكافئة للفرضية الصفرية. ولننظر مثلاً إلى الفرضية الصفرية التي تفيد بأن الجودة السمعية المدركة في المجتمع الإحصائي متماثلة لجميع الظروف. ويكافئ ذلك حساب تناقضات عمودية عددها $K - 1$، مثلاً عن طريق تكوين متغيرات فروق بين الظروف البالغ عددها K، ثم اختبار فرضية أن المتجه μ الذي يتكون من متوسط المجتمع الإحصائي لجميع التناقضات البالغ عددها $K - 1$ يساوي المتجه الصفري، أي $\mu = 0$. وعلى سبيل المثال، إذا عُرض المرجع ومشفران، يمكن إنشاء مُتَغَيّري للفروق D_1 وD_2 عن طريق حساب الفرق بين تقييم المرجع وتقييم المشفر A (D_1) لكل مقيّم والفرق بين تقييم المشفر A وتقييم المشفر B (D_2). ويستند تحليل التباين بالقياسات المتكررة الذي يستعمل النهج متعدد المتغيرات إلى متغيرات الفروق ويستعمل اختباراً متعدد المتغيرات لفرضية أن $\mu = 0$. وفي هذا النهج لا تكون هناك حاجة إلى افتراضات بشأن مصفوفة التباين–التباين المشترك. وبالنسبة للبيانات التي تتبع توزيعاً طبيعياً متعدد المتغيرات، يعتبر هذا الاختبار اختباراً دقيقاً، ولكنه يتطلب أن يكون عدد المقيّمين مساوياً على الأقل لعدد مستويات العوامل. ولذلك، لا يمكن استعماله إذا عُرِضت 9 ظروف (8 مشفرات بالإضافة إلى المرجع) مثلاً على 8 مقيّمين فقط.

وتعتمد القدرة النسبية للنهجين على العديد من العوامل من بينها حجم العينة وعدد مستويات العوامل المتضمنة في عامل داخل مشاهدات العينة. ووفقاً لدراسة Algina and Keselman (1997)، تتمثل قاعدة اختيار بسيطة في استعمال النهج أحادي المتغيرات مع تصحيح هيونه–فلدت إذا كانت $\tilde{\varepsilon} > 0{,}85$ و$N < K + 30$ حيث N هي عدد المقيّمين، وK هي الحد الأقصى لعدد مستويات العوامل داخل مشاهدات العينة. وفي الحالات المتبقية، ينبغي استعمال النهج متعدد المتغيرات. وتجدر ملاحظة أنه إذا أُجريت التجربة في مختبرات مختلفة، فإن N ستكون العدد الإجمالي للمقيّمين المشاركين في الدراسة (مثلاً، 10 مقيّمين في المختبر A و10 مقيّمين في المختبر B، يعني أن $N = 20$).

4 إجراء تحليل التباين بالقياسات المتكررة والاختبارات اللاحقة الاختيارية

في هذه الخطوة، تُجرى اختبارات شاملة لتأثيرات الظرف والمادة السمعية وتفاعلهما باستعمال أسلوب تحليل التباين بالقياسات المتكررة. ولحساب تحليل التباين بالقياسات المتكررة، تتطلب معظم حزم البرمجيات مثل SAS، وSPSS، وStatistica، توافر البيانات في شكل "صف واحد لكل مقيّم". وبالتالي، يجب أن يحتوي جدول البيانات على صف واحد فقط لكل مقيّم، وتُعرض تقييمات كل تركيبات الظرف والمادة السمعية في شكل أعمدة ("متغيرات").

ويوفر تحليل التباين بالقياسات المتكررة ذو العاملين معلومات عن ثلاثة تأثيرات.

1) التأثير الرئيسي للظرف

في معظم الحالات، سيكون هذا هو اختبار الاهتمام الرئيسي. فإذا أشار تحليل التباين إلى وجود تأثير دال للظرف، يمكن رفض الفرضية الصفرية التي تفيد بتماثل جودة المادة السمعية المدركة بالنسبة لكل الظروف في المجتمع الإحصائي (المرجع، والمشفر 1 إلى k). وبعبارة أخرى، فإن الاختبار يشير إلى أن هناك فروقاً بين الجودة السمعية المدركة للأنظمة السمعية في المجتمع الإحصائي. وكقياس

لحجم التأثير، لا يمكن استعمال d كوهين (Cohen) [6]، أو واحد من نظرائها، لأن d غير محددة لمقارنة أكثر من متوسطين. ومن الشائع في سياق تحليل التباين الإبلاغ عن قياس لقوة الارتباط. وتوفر هذه القياسات معلومات عن نسبة التباين في البيانات التي يمكن أن تعزى إلى التأثير قيد النظر. وهو نفس الأساس المنطقي الكامن وراء مُعامِل التحديد، R^2. ويمكن لمعظم حزم البرمجيات الإحصائية حساب η^2 الجزئية التي تُحسب بوصفها نسبة التباين الناتج عن التأثير إلى مجموع تباين التأثير وتباين الخطأ (القيمة المتبقية). ويمكن الاطلاع على مناقشة للقياسات البديلة لقوة الارتباط في دراسة Olejnik and Algina [31].

وبعد الوصول إلى نتيجة اختبار دالة لتأثير رئيسي، سيكون من المهم في كثير من الأحيان تحديد موقع منشأ هذا التأثير. ويمكن تحقيق ذلك عن طريق حساب تناقضات محددة. وعلى سبيل المثال، قد يهتم المرء بما إذا كانت جودة الصوت في مشفر جديد تختلف عن جودة صوت ثلاثة أنظمة قائمة. وللإجابة على هذا السؤال، يحسب المرء أولاً التقييم المتوسط للمشفرات الثلاثة القائمة لكل مقيّم، ويحصل على المتوسط عبر المادة السمعية. ونتيجة لذلك، سيكون هناك لكل مقيّم (أ) تقييم واحد للمشفر الجديد، و(ب) تقييم متوسط للمشفرات الثلاثة الأخرى. ثم تُقارن هاتان القيمتان باستعمال اختبار t لأزواج العينات. ويلاحَظ أنه نظراً إلى أن البيانات مستمدة من تصميم للقياسات المتكررة، فمن المهم عدم استعمال التباين المجمع [27]. ويُلاحَظ أيضاً أن هذا التناقض ربما جرى اختباره كتناقض مخطط بدلاً من إجراء تحليل التباين. ويوصَى بشكل عام باستعمال اختبارات الدلالة ذات الذيلين. غير أنه إذا كان هناك مثلاً فرضية بديهية تفيد بأن المشفر الجديد ينبغي أن يحصل على تقييمات أفضل من المشفرات القائمة، فمن المسموح به في هذه الحالة استعمال منطقة رفض ذات ذيل واحد.

ويمكن حساب تناقضات محددة أخرى باستعمال نفس الأساس المنطقي. وهناك صيغة أكثر عمومية لاختبار التناقضات تتمثل في حساب تركيب خطي من التقييمات التي يتم الحصول عليها في الظروف التجريبية المختلفة، ثم استعمال اختبار t بعينة واحدة لتقرير ما إذا كان هذا التناقض يختلف اختلافاً دالاً عن 0. ولكل مقيّم i تُحسب قيمة تناقض

$$\Psi_i = \sum_{j=1}^{a} c_j Y_{ij}, \quad \sum_{j=1}^{a} c_j = 0,$$

حيث Y_{ij} هي التقييم الذي يضعه المقيّم i في الظرف j (المتوسط عبر المادة السمعية)، وa هي عدد الظروف التي وضعت في الاعتبار في هذا التناقض، وc_j هي المعاملات. وفي المثال أعلاه، إذا كان المشفر الجديد يقابل $j=1$ والمشفرات الثلاثة الأخرى تقابل $j = 2...4$، فإن اختيار $c_1 = -1$ و$c_2 = c_3 = c_4 = 1/3$ سيعطينا اختباراً لفرضية أن الجودة السمعية في المشفر الجديد تختلف عن المشفرات الثلاثة الأخرى.

وإذا حُسب أكثر من تناقض لاحق واحد، فإن هذا كما نوقش أعلاه سيؤدي إلى مشاكل في الاختبار المتعدد. ولحل هذه المشكلة، يوصَى بتطبيق إجراء بونفيروني المعزز المقبول تسلسلياً الذي أعده هوشبرغ [15]. ويتحكم هذا الإجراء في معدل الخطأ من النوع الأول، في حين أنه أكثر قوة من العديد من الإجراءات البديلة [20]. وفي إجراء هوشبرغ، يقوم المرء أولاً بحساب تناقضات الاهتمام m ويرتبها وفقاً للقيمة p. ثم يبدأ بفحص أكبر قيمة من قيم p. وإذا كانت قيمة p هذه أصغر من α، ترفض كل الفرضيات (أي أن كل التناقضات دالة). وإذا لم يكن الأمر كذلك، فإن اختبار t الذي استعملت فيه أكبر قيمة من قيم p لم يكن دالاً، فيستمر المرء في الحساب لمقارنة قيمة p التالية الأصغر بقيمة $\alpha/2$. وإذا كانت هذه القيمة أصغر، فإن هذا الاختبار وجميع الاختبارات التي تكون قيمة p فيها أصغر ستكون دالة. وإذا لم يكن الأمر كذلك، فإن الاختبار الذي استعملت فيه ثاني أكبر قيمة من قيم p لم يكن دالاً، فيمضي المرء لمقارنة قيمة p الأصغر التالية بقيمة $\alpha/3$. وبصورة أكثر رسمية، إذا كانت p_i و$m = i$، $m - 1$، ...، 1 هي قيم p في نظام تنازلي، ففي حالة أي $m = i$، $m - 1$، ...، 1، إذا كانت $p_i < \alpha/(m - i + 1)$ فإن كل الاختبارات التي تكون فيها $i' \leq i$ ستكون دالة.

ومن حيث المبدأ، يمكن أيضاً حساب المقارنات الزوجية اللاحقة بين التقييمات لكل الظروف. وبالنسبة لتصميم القياسات المتكررة، سيتطلب ذلك حساب اختبارات t للعينات الزوجية بين جميع أزواج الظروف. غير أنه لا يوصَى بهذا النهج. ولننظر مثلاً إلى تجربة تُستعمل فيها 7 مشفرات ومرجع واحد. فبالنسبة لهذه المجموعة التي تتكون من 8 ظروف، يمكن حساب 28 اختباراً من الاختبارات الزوجية ($28 = 7/2 \cdot 8$)، ولن يكون من السهل استخلاص معلومات ذات مغزى من هذا العدد الكبير من الاختبارات. وإذا تم

اختبار جميع الفوارق الزوجية، فإن العدد الكبير من الاختبارات التي تطبق إجراء هوشبرغ [15] لتصحيح الاختبار المتعدد سيكون بالطبع ذا أهمية خاصة. ويلاحَظ أنه إذا كانت هناك أدلة على حدوث انحراف عن التوزيع الطبيعي لدرجات الفرق التي يقوم عليها اختبار *t* للعينات الزوجية، فإن اختباراً بديلاً لا يفترض وجود التوزيع الطبيعي هو اختبار العلامة.

وينبغي ملاحظة أنه بعد حدوث تأثير رئيسي دال، يمكن ألا يكون أي من التناقضات اللاحقة أو الفوارق الزوجية دالاً [28]، بسبب المعلومات الإحصائية المختلفة المستعملة في تحليل التباين بالقياسات المتكررة والاختبارات اللاحقة. والمهم هو أن تحليل التباين بالقياسات المتكررة هو الاختبار الأنسب. ولذلك، فإن التأثير الدال الذي يشير إليه تحليل التباين يظل صالحاً حتى إذا لم يكن أي من الاختيارات اللاحقة دالاً. وإذا لم يكن أي تناقض لاحق دالاً بعد اختبار دلالة شامل (ANOVA)، فإنه يمكن الخلوص إلى أن الأنظمة السمعية تختلف من حيث جودة الصوت المدركة. ويمكن أيضاً مقارنة الفوارق بين الأنظمة السمعية ببعضها البعض. فمثلاً، بالنسبة لأزواج الأنظمة السمعية التي يظهر فيها أعلى اختلاف في تقييمات جودة الصوت، يحتمل أن تصبح هذه الاختلافات الزوجية ذات دلالة مع حجم عينة أكبر. غير أنه يجب أن نخلص إلى أنه لا توجد فوارق زوجية ذات دلالة في الدراسة الحالية.

وإذا *لم يُظهر* تحليل التباين بالقياسات المتكررة أي تأثير رئيسي دال للظرف، فإن هذا يشير إلى أن الفوارق بين الأنظمة التي يجري اختبارها كانت صغيرة. غير أنه، وبسبب حجم العينة المحدود، لا يمكن أن نخلص إلى أنه *لا توجد* فوارق في المجتمع الإحصائي من حيث الجودة السمعية المدركة بين الظروف [3]. ويمكن أن تكون فوارق المجتمع الإحصائي إما صفراً أو أن تكون أحجام التأثيرات أصغر من أن تُكتشف نظراً لحجم العينة. وإذا أُجري تحليل مسبق للقدرة، أي أنه قد تم اختيار حجم العينة بحيث تكون كافية لاكتشاف حجم تأثير محدد باحتمال محدد، فيمكن أن نخلص إلى أن البيانات تعتبر دليلاً على عدم وجود تأثير للحجم المحدد سلفاً.

ويمكن اعتبار هذه النتيجة تعريفاً لشفافية المشفرات. وإذا لم يتم إجراء أي تحليل مسبق للقدرة، يتعين توخي الحذر عند الخلوص إلى أن المشفرات شفافة، للأسباب التي سبق شرحها. ومن الحلول التقريبية المعتادة اللاحقة هي مقارنة قيمة *p* بقيمة [0,2] وليس 0,05. وإذا ظل الاختبار غير دال فإن هذا يعتبر دليلاً أقوى إلى حد ما عن عدم وجود فوارق في الجودة السمعية المدركة للظروف.

2) *التأثير الرئيسي للمادة السمعية*

باستعمال نفس الخطوات ونفس الأساس المنطقي المستعمل أعلاه، فإن اختبار التأثير الرئيسي للمادة السمعية يوفر معلومات عن التغيرات النظامية في التقييمات وفقاً لمادة الاختبار. وبالنسبة لمعظم سيناريوهات اختبار MUSHRA، فإن هذا التأثير لا ينبغي أن يكون ذا أهمية كبيرة لأنه لا يتعلق بالفرق بين الأنظمة السمعية.

3) *تفاعل الظرف والمادة السمعية*

إذا أظهر تحليل التباين بالقياسات المتكررة تفاعلاً دالاً للظرف والمادة السمعية، فإن تأثير النظام السمعي على الجودة السمعية المدركة يختلف بين مواد الاختبار. وعلى سبيل المثال، يمكن تقييم المرجع وأحد المشفرات بشكل متساو لأغنية شعبية جرى ضغطها بدرجة كبيرة وتم فيها حجب الأصوات المصطنعة للتشفير بمكونات تشويش موجودة في المادة. ومن الناحية أخرى، يمكن أن يكون تقييم جودة الصوت للمشفر أقل من المرجع في تسجيل يتميز بمدى دينامي مرتفع لحفل موسيقى كبير. وعادة ما سيكون هذا التفاعل ذا أهمية في اختبار MUSHRA لأنه يشير إلى أن الفرق بين الأنظمة السمعية يعتمد على مادة الاختبار.

وبعد إجراء اختبار شامل دال لتأثير التفاعل، يمكن مواصلة استكشاف طبيعة التفاعل باستعمال الاختبارات اللاحقة. ومن النُهُج الشائعة اختبار لما تسمى *التأثيرات الرئيسية البسيطة*. ويمكن حساب هذه مثلاً عن طريق إجراء عدة تحليلات تباين بالقياسات المتكررة تكون منفصلة وذات عامل واحد باستعمال ظرف عامل بين مشاهدات العينة، بحيث يُنفّذ تحليل واحد لكل مادة سمعية. وستوضح هذه التحليلات المواد السمعية التي كان للظرف فيها تأثير دال. ومرة أخرى، ينبغي استعمال إجراء هوشبرغ كتصحيح للاختبار المتعدد.

وكما أشير أعلاه، فإن جميع الفوارق الزوجية بين تركيبات الظرف والمادة السمعية يمكن اختبارها من حيث المبدأ باستعمال اختبارات *t* منفصلة للعينات الزوجية وإجراء هوشبرغ. وسيكون عدد المقارنات الزوجية أعلى من عدد مقارنات التأثيرات الرئيسية. غير أنه إذا تم مزج 8 أنظمة سمعية مثلاً بأربع مواد اختبار، فستكون هناك 24 تركيبة من الأنظمة السمعية ومادة الاختبار بما يقابل 276 (276 = 23/2 · 24) اختباراً من الاختبارات الزوجية. ومن الواضح أنه لا يمكن التوصية باستعمال هذا النهج.

5 التوسع إلى تصميمات تحتوى على متغير (تجميع) بين مشاهدات العينة

نظرنا حتى الآن في تصميم بدون عوامل ما بين مشاهدات العينة. فأي تحليلات يمكن إجراؤها إذا أُجري الاختبار على مجموعات مختلفة من المقيّمين، مثلاً في مختبرين، أو للموسيقيين مقابل غير الموسيقيين؟

وإذا كانت هناك عوامل قائمة بين مشاهدات العينة، فمما له أهمية بالغة تقرير ما إذا كان عدد المقيّمين متماثلاً في كل المجموعات (التصميم المتوازن) أو مختلفاً بين المجموعات (التصميم غير المتوازن).

التصميم المتوازن. إذا كان عدد المقيّمين متماثلاً لجميع مستويات عامل بين مشاهدات العينة، أو إذا كانت أحجام المجموعات لا تختلف بنسب تزيد عن 10 في المائة، فإنه يمكن استعمال النهج أحادي المتغيرات مع تصحيح هيونه–فلدت لدرجات الحرية أو النهج متعدد المتغيرات لإجراء تحليل التباين بالقياسات المتكررة [21]. وسيحتوي التصميم الآن على ظرف عوامل داخل مشاهدات العينة والمادة السمعية وعلى الأقل عامل واحد من عوامل ما بين مشاهدات العينة (مثلاً المختبر). ولذلك، سيوفر تحليل التباين بالقياسات المتكررة اختباراً إضافياً للتأثير (التأثيرات) ما بين مشاهدات العينة فضلاً عن التفاعلات بين التأثيرات داخل مشاهدات العينة وفيما بينها.

وقد يتضح مثلاً أن تفاعل الظرف والمختبر هو تفاعل دال، وهو ما يعني أن الفوارق في الجودة السمعية المدركة للأنظمة السمعية تختلف من المختبر A إلى المختبر B. ويلاحَظ أننا نفترض هنا أن نفس التركيبات للظرف والمادة السمعية عُرضت على جميع المجموعات. وإذا عرضت مواد سمعية مختلفة مثلاً في المختبرين فإنه لا يمكن استعمال الطرائق المقترحة هنا. وبدلاً من ذلك، فإن ما يطلق عليها نماذج التأثيرات العشوائية ستكون هي المطلوبة [28]، وهي خارج نطاق هذا المرفق.

التصميم غير المتوازن. إذا اختلفت أحجام المجموعات بنسب تزيد عن 10 في المائة، فإن النهج أحادي المتغيرات والنهج متعدد المتغيرات لن يعطيانا للأسف أي نتائج صالحة للاختبارات [21]. ولذلك يوصَى بشدة بالتخطيط لأحجام مجموعات متساوية وبذلك يتم تجنب هذه المشكلة. وإذا كانت أحجام المجموعات غير متساوية، يمكن التوصية بإجراءين للتحليل. والنهج الأول هو اختبار التقريب العام (IGA) المحسن [1]، والنهج الثاني هو أسلوب معين من أساليب طريقة الاحتمالات القصوى القائمة على تحليل النماذج المختلطة [23]. ويتوفر اختبار التقريب العام المحسن كتعليمات تشغيل في برنامج SAS. ويمكن إجراء تحليل النموذج المختلط مثلاً في SAS PROC MIXED. وبالنسبة لهذا التحليل الأخير، هناك خياران مهمان. والخيار الأول هو أنه يجب حساب درجات الحرية وفقاً للطريقة الواردة في [19]، التي تتحقق في برنامج SAS بوضع الخيار ddfm=KR في بيان خاصية النموذج (model). والخيار الثاني هو أنه يجب وضع بنية التغاير المشترك غير المنظم (UN-H) بين مشاهدات العينة [23]، وذلك باستعمال خيارات type=UN group=groupingvar في البيان المتكرر، حيث groupingvar هو اسم المتغير الذي يحتوي على تصنيف المجموعة.

المراجع

[1] Algina, J. (1997). Generalization of Improved General Approximation tests to split-plot designs with multiple between-subjects factors and/or multiple within-subjects factors. *British Journal of Mathematical and Statistical Psychology, 50,(2),* 243-252.

[2] Algina, J., & Keselman, H. J. (1997). Detecting repeated measures effects with univariate and multivariate statistics. *Psychological Methods, 2*(2), 208-218.

[3] Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal, 311*(7003), 485-485.

[4] Arnau, J., Bendayan, R., Blanca, M. J., & Bono, R. (2013). The effect of skewness and kurtosis on the robustness of linear mixed models. *Behavior Research Methods, 45*(3), 873-879. doi: 10.3758/s13428-012-0306-x.

[5] Berkovits, I., Hancock, G. R., & Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: relative robustness to sphericity and normality violations. *Educational and Psychological Measurement, 60*(6), 877-892.

[6] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.

[7] Conover, W. J. (1999). Practical nonparametric statistics (3rd ed.). New York: Wiley.

[8] Cramér, H. (1946). Mathematical methods of statistics. Princeton: Princeton University Press.

[9] DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*(3), 292-307. doi: 10.1037//1082-989x.2.3.292.

[10] Doornik, J. A., & Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics, 70,(s1),* 927-939. doi: 10.1111/j.1468-0084.2008.00537.x.

[11] Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288. doi: 10.3102/00346543042003237.

[12] Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*(2), 95-112.

[13] Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte-Carlo results in methodological research: The one-factor and two-factor fixed effects ANOVA cases. *Journal of Educational and Behavioral Statistics, 17*(4), 315-339. doi: 10.3102/10769986017004315.

[14] Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods, 19*(10), 3595-3617. doi: 10.1080/03610929008830400.

[15] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800-802.

[16] Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact *F*-distributions. *Journal of the American Statistical Association, 65*(332), 1582-1589.

[17] Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational aand Behavioral Statistics, 1*(1), 69-82. doi: http://dx..org/10.2307/1164736.

[18] Jensen, D. R. (1982). Efficiency and robustness in the use of repeated measurements. *Biometrics, 38*(3), 813-825. doi: 10.2307/2530060.

[19] Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*(3), 983-997.

[20] Keselman, H. J. (1994). Stepwise and simultaneous multiple comparison procedures of repeated measures' means. *Journal of Educational and Behavioral Statistics, 19*(2), 127-162.

[21] Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical & Statistical Psychology, 54, (1),* 1-20.

[22] Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. M., & Wilcox, R. R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical & Statistical Psychology, 53,(2),* 175-191.

[23] Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted *F* tests. *Educational and Psychological Measurement, 64*(2), 224-242. doi: 10.1177/0013164403260196.

[24] Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, N.C.: SAS Institute, Inc.

[25] Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*(4), 579-619. doi: 10.2307/1170654.

[26] Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519-530. doi: 10.2307/2334770.

[27] Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. *Journal of Educational and Behavioral Statistics, 5*(3), 269-287. doi: 10.3102/10769986005003269.

[28] Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates.

[29] Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.

[30] Oberfeld, D., & Franke, T. (2013). Evaluating the robustness of repeated measures analyses: The case of small sample sizes and non-normal data. *Behavior Research Methods, 45*(3), 792-812. doi: http:/dx.doi.org/10.3758/s13428-012-0281-2.

[31] Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434-447. doi: 10.1037/1082-989x.8.4.434.

[32] Rasmussen, J. L. (1987). Parametric and Bootstrap Approaches to Repeated Measures Designs. *Behavior Research Methods Instruments & Computers, 19*(4), 357-360.

[33] Rouanet, H., & Lépine, D. (1970). Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology, 23*(2), 147-163.

[34] Royston, J. P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk-W. *Applied Statistics-Journal of the Royal Statistical Society Series C, 32*(2), 121-133. doi: 10.2307/2347291.

[35] Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology-European Journal of Research Methods for the Behavioral and Social Sciences, 6*(4), 147-151. doi: 10.1027/1614-2241/a000016.

[36] St. Laurent, R., & Turk, P. (2013). The effects of misconceptions on the properties of Friedman's test. *Communications in Statistics-Simulation and Computation, 42*(7), 1596-1615. doi: 10.1080/03610918.2012.671874.

[37] Tukey, J. W. (1977). Exploratory data analysis. Reading, Mass.: Addison-Wesley Pub. Co.

[38] Seco, G. V., Izquierdo, M. C., García, M. P. F., & Díez, F. J. H. (2006). A comparison of the bootstrap-F, improved general approximation, and Brown-Forsythe multivariate approaches in a mixed repeated measures design. *Educational and Psychological Measurement, 66*(1), 35-62.

[39] Wilcox, R. R., Keselman, H. J., Muska, J., & Cribbie, R. (2000). Repeated measures ANOVA: Some new results on comparing trimmed means and means. *British Journal of Mathematical & Statistical Psychology, 53*, 69-82.

المرفق 5
بالملحق 1
(إعلامي)

متطلبات السلوك الأمثل للمرتكزات

ترد فيما يلي الواصفات الرئيسية التي يجب أن تتوفر في أي مرتكز ناجح عند تصميمه.

وعلى السلوك الأمثل للمرتكز أن:

1) ينتج بيانات لا تظهر أي تغيرات كبيرة في الترتيبات النسبية لأنظمة الاختبار عند مقارنتها بالبيانات التي جمعت باستعمال مواصفات المرتكزات الواردة في التوصية ITU-R BS.1534؛

2) يرتبط بتقييمات المستمعين التي تستعمل مدى أوسع في مقياس تقييم أنظمة الاختبار عند مقارنته بالبيانات التي جمعت للأنظمة الخاضعة للاختبار باستعمال مواصفات المرتكزات الواردة في التوصية ITU-R BS.1534؛

3) يدركه المستمعون بوصفه أكثر شبهاً بأنظمة الاختبار منه بالمرتكزات المشار إليها في المواصفات الواردة في التوصية ITU-R BS.1534. وقد يؤدي هذا بدوره إلى أوقات تقييم أطول للمرتكزات؛

4) يسمح بإجراء مقارنة حساسة لأنظمة الاختبار متوسطة المدى؛

5) ينتج درجات تختلف بين المرتكز منخفض المدى والمرتكز متوسط المدى بحوالي 20-30 نقطة؛

6) ينتج انحطاطاً في الجودة في المرتكزات التي تعتمد بشكل محدود على المحتوى.