Международный союз электросвязи

МСЭ-R
Сектор радиосвязи МСЭ

Рекомендация МСЭ-R BS.1534-2
(06/2014)

Метод субъективной оценки промежуточного уровня качества аудиосистем

Серия BS
Радиовещательная служба (звуковая)

ITU Международный союз электросвязи

Предисловие

Роль Сектора радиосвязи заключается в обеспечении рационального, справедливого, эффективного и экономичного использования радиочастотного спектра всеми службами радиосвязи, включая спутниковые службы, и проведении в неограниченном частотном диапазоне исследований, на основании которых принимаются Рекомендации.

Всемирные и региональные конференции радиосвязи и ассамблеи радиосвязи при поддержке исследовательских комиссий выполняют регламентарную и политическую функции Сектора радиосвязи.

Политика в области прав интеллектуальной собственности (ПИС)

Политика МСЭ-R в области ПИС излагается в общей патентной политике МСЭ-T/МСЭ-R/ИСО/МЭК, упоминаемой в Приложении 1 к Резолюции МСЭ-R 1. Формы, которые владельцам патентов следует использовать для представления патентных заявлений и деклараций о лицензировании, представлены по адресу http://www.itu.int/ITU-R/go/patents/en, где также содержатся Руководящие принципы по выполнению общей патентной политики МСЭ-T/МСЭ-R/ИСО/МЭК и база данных патентной информации МСЭ-R.

Серии Рекомендаций МСЭ-R

(Представлены также в онлайновой форме по адресу http://www.itu.int/publ/R-REC/en)

Серия	Название
BO	Спутниковое радиовещание
BR	Запись для производства, архивирования и воспроизведения; пленки для телевидения
BS	**Радиовещательная служба (звуковая)**
BT	Радиовещательная служба (телевизионная)
F	Фиксированная служба
M	Подвижные службы, служба радиоопределения, любительская служба и относящиеся к ним спутниковые службы
P	Распространение радиоволн
RA	Радиоастрономия
RS	Системы дистанционного зондирования
S	Фиксированная спутниковая служба
SA	Космические применения и метеорология
SF	Совместное использование частот и координация между системами фиксированной спутниковой службы и фиксированной службы
SM	Управление использованием спектра
SNG	Спутниковый сбор новостей
TF	Передача сигналов времени и эталонных частот
V	Словарь и связанные с ним вопросы

Примечание. *– Настоящая Рекомендация МСЭ-R утверждена на английском языке в соответствии с процедурой, изложенной в Резолюции МСЭ-R 1.*

Электронная публикация
Женева, 2015 г.

РЕКОМЕНДАЦИЯ МСЭ-R BS.1534-2

Метод субъективной оценки промежуточного уровня качества аудиосистем

(Вопрос МСЭ-R 62/6)

(2001-2003-2014)

Сфера применения

В настоящей Рекомендации описывается метод субъективной оценки промежуточных уровней качества звучания. В нем воспроизводятся многие аспекты Рекомендации МСЭ-R BS.1116 и используется та же шкала, что и для оценки качества изображений (см. Рекомендацию МСЭ-R BT.500).

Этот метод, известный под названием "тест при использовании нескольких входных сигналов со скрытым эталонным сигналом и с опорной точкой" (MUSHRA), успешно прошел испытания. В ходе испытаний было продемонстрировано, что метод MUSHRA пригоден для оценки промежуточного качества звучания и дает точные и надежные результаты.

Ключевые слова

Испытание с прослушиванием, артефакты, промежуточное качество звука, кодирование звуковых сигналов, субъективная оценка, качество звука.

Ассамблея радиосвязи МСЭ,

учитывая,

a) что Рекомендациями МСЭ-R BS.1116, МСЭ-R BS.1284, МСЭ-R BT.500, МСЭ-R BT.710 и МСЭ-R BT.811, а также Рекомендациями МСЭ-T P.800, МСЭ-T P.810 и МСЭ-T P.830 установлены методы субъективной оценки качества аудио-, видео- и речевых систем;

b) что новые виды служб доставки контента, такие как потоковое аудио в интернете и на твердотельных проигрывателях, цифровые спутниковые службы, цифровые системы коротковолнового и средневолнового радиовещания, а также подвижные мультимедийные системы могут работать при промежуточном качестве звука;

c) что Рекомендация МСЭ-R BS.1116 предназначена для оценки небольших ухудшений качества и непригодна для оценки систем с промежуточным качеством звука;

d) что Рекомендация МСЭ-R BS.1284 не предусматривает абсолютной шкалы для оценки промежуточного качества звука;

e) что наличие надлежащих и релевантных опорных сигналов в ходе испытаний создает условия для устойчивого использования шкалы субъективной оценки;

f) что Рекомендации МСЭ-T P.800, МСЭ-T P.810 и МСЭ-T P.830 предназначены главным образом для субъективной оценки речевых сигналов в телефонии и оказались на практике недостаточными для оценки звуковых сигналов в сфере радиовещания;

g) что использование стандартизированных методов субъективной оценки важно для обмена данными испытаний, а также для обеспечения совместимости и правильной оценки таких данных;

h) что для новых мультимедийных служб может требоваться совместная оценка качества звука и изображения;

i) что наименование MUSHRA зачастую безосновательно употребляется в отношении испытаний, в которых не применяются эталонный сигнал и опорные сигналы;

j) что опорные точки могут влиять на результаты испытаний и желательным их качеством является сходство с артефактами испытуемой системы,

рекомендует

1 использовать методики испытания и оценки, изложенные в Приложении 1 к настоящей Рекомендации, для субъективной оценки промежуточного качества звучания,

далее рекомендует

1 продолжить исследования опорных сигналов, обладающих характеристиками ухудшений, которые встречаются в современных аудиосистемах, и обновлять настоящую Рекомендацию с включением в нее новых опорных сигналов по мере целесообразности.

Приложение 1

1 Введение

В настоящей Рекомендации описывается метод субъективной оценки промежуточных уровней качества звучания. В нем воспроизводятся многие аспекты Рекомендации МСЭ-R BS.1116 и используется та же шкала, что и для оценки качества изображений (см. Рекомендацию МСЭ-R BT.500).

Этот метод, известный под названием "тест при использовании нескольких входных сигналов со скрытым эталонным сигналом и с опорной точкой" (MUSHRA), успешно прошел испытания. В ходе испытаний было продемонстрировано, что метод MUSHRA пригоден для оценки промежуточного качества звучания и дает точные и надежные результаты ([2], [4], [3]).

Настоящая Рекомендация содержит следующие разделы и приложения:

Раздел 1. Введение

Раздел 2. Сфера применения, обоснование целесообразности и назначение нового метода

Раздел 3. План эксперимента

Раздел 4. Отбор оценщиков

Раздел 5. Метод испытания

Раздел 6. Атрибуты

Раздел 7. Тестовый материал

Раздел 8. Условия прослушивания

Раздел 9. Статистический анализ

Раздел 10. Протокол и представление результатов испытаний

Приложение 1 (нормативное). Инструкция оценщикам

Приложение 2 (информативное). Руководящие указания по проектированию пользовательских интерфейсов

Приложение 3 (нормативное). Описание непараметрического статистического сравнения двух образцов с использованием методов перевыборки и численных методов Монте-Карло

Приложение 4 (информативное). Руководящие указания по параметрическому статистическому анализу

Приложение 5 (информативное). Требования к оптимальному поведению опорных сигналов

2 Сфера применения, обоснование целесообразности и назначение нового метода

Субъективная оценка с прослушиванием до сих пор считается самым надежным способом определения качества аудиосистем. Существуют подробно описанные и проверенные на практике методы оценки качества звука в верхнем и нижнем диапазонах шкалы качества.

Для оценки высококачественных звуковых систем с небольшими ухудшениями качества используется Рекомендация МСЭ-R BS.1116 "Методы субъективной оценки небольшого ухудшения качества в звуковых системах", включая многоканальные звуковые системы. Однако есть области применения, в которых снижение качества звука допустимо или неизбежно. Стремительный рост популярности интернета как среды для распространения аудиоматериалов и звукового вещания в условиях ограниченной скорости передачи данных заставил поступиться качеством звука. Другие области, в которых можно столкнуться с промежуточным качеством звука, – это цифровое АМ-радиовещание (в частности, Всемирное цифровое радио, или DRM), цифровое спутниковое радиовещание, комментаторские каналы на радио и телевидении, услуги аудио по запросу и передача звука по телефонным линиям. Метод испытания, определенный в Рекомендации МСЭ-R BS.1116, не вполне подходит для оценки этих более низкокачественных аудиосистем [4] ввиду недостаточной способности к выявлению незначительных различий по качеству в нижней части шкалы.

Методы, содержащиеся в Рекомендации МСЭ-R BS.1284, ориентированы либо на высококачественный звук, либо не предусматривают абсолютной шкалы качества.

Другие Рекомендации, например МСЭ-Т P.800, МСЭ-Т P.810 и МСЭ-Т P.830, предназначены главным образом для субъективной оценки речевых сигналов в телефонии. Группа по проекту B/AIM Европейского радиовещательного союза (EPC) провела эксперименты по оценке качества звука указанными выше методами МСЭ-Т на типичных для радиовещательных систем аудиоматериалах. Ни один из этих методов не отвечает одновременно следующим требованиям: наличие абсолютной шкалы, сравнение с эталонным сигналом, узкие доверительные интервалы и разумное количество оценщиков. Поэтому ни один из этих методов не позволяет надлежащим образом оценить качество звукового сигнала в радиовещательной среде.

Пересмотренный метод испытания, описанный в настоящей Рекомендации, призван обеспечить надежные и воспроизводимые измерения для систем с качеством звучания, которое обычно попадает в нижнюю половину шкалы ухудшений, применяемой в Рекомендации МСЭ-R BS.1116 ([2], [4], [3]). В методе испытания MUSHRA используется высококачественный эталонный сигнал и предполагается, что испытуемые системы вносят значительные ухудшения. Метод MUSHRA предназначен для оценки аудиосистем промежуточного качества. Если метод MUSHRA используется с надлежащим контентом, то в идеальном случае данные слушателями оценки должны попадать в диапазон от 20 до 80 баллов по шкале MUSHRA. Если же оценки для большинства условий испытаний находятся в диапазоне от 80 до 100 баллов, результаты испытаний могут быть недействительны.

Вероятные причины суженного и смещенного диапазона оценок – привлечение неопытных оценщиков, использование некритичного контента или ненадлежащий выбор метода испытания для испытуемых алгоритмов кодирования.

3 План эксперимента

Для сбора надежной информации в представляющих интерес областях научного знания применяется множество различных исследовательских стратегий. Для субъективной оценки ухудшений качества в аудиосистемах применяются наиболее формальные экспериментальные методы. Субъективные эксперименты характеризуются прежде всего реальным контролем над условиями эксперимента и варьированием этих условий, а кроме того сбором и анализом статистических данных, полученных от слушателей. В целях сведения к минимуму неконтролируемых факторов, которые могут привести к неоднозначности экспериментальных результатов, необходимо тщательное планирование экспериментов. Например, если фактическая последовательность звуковых элементов идентична для всех оценщиков, то нельзя с уверенностью сказать, чем были обусловлены вынесенные ими суждения – самой этой последовательностью или разными уровнями ухудшений, представленных вниманию оценщиков. Соответственно, схема представления условий испытания должна быть такой, чтобы обнаруживать влияние независимых факторов и только этих факторов.

Если есть основания ожидать, что потенциальные ухудшения и другие характеристики будут распределены равномерно на протяжении всего испытания, то условия испытания можно представлять в истинно рандомизированном порядке. Если же предполагается наличие неоднородности, ее необходимо учитывать при представлении условий испытания. Например, если оцениваемый материал неоднороден по уровню трудности, порядок представления входных сигналов должен быть распределен случайно как в пределах одного сеанса, так и между сеансами.

Испытания с прослушиванием необходимо планировать так, чтобы качество выносимых оценщиками суждений не снижалось из-за перегрузки. Предпочтительно производить оценку аудиосистем без сопровождающих изображений, за исключением случаев, когда соотношение между звуком и изображением является значимым. Немаловажно предусмотреть надлежащие контрольные условия. Такие условия обычно включают представление аудиоматериалов неухудшенного качества способами, непредсказуемыми для оценщиков. Именно различия в суждениях о контрольных и потенциально ухудшенных входных сигналах и дают основания заключить, что поставленные оценки действительно характеризуют ухудшения.

Некоторые из этих соображений будут рассмотрены позже. Следует понимать, что планирование и выполнение экспериментов, а также статистический анализ их результатов – сложные темы, не все аспекты которых могут быть изложены в Рекомендации наподобие этой. Рекомендуется проконсультироваться с профессионалами в сфере планирования экспериментов и статистики или привлечь их к работе на начальной стадии планирования испытания с прослушиванием.

В целях создания условий для эффективного анализа данных и переноса их между лабораториями необходимо включить план эксперимента в протокол. В плане следует детально определить как зависимые, так и независимые переменные. Число независимых переменных определяется вместе с соответствующими уровнями.

4 Отбор оценщиков

Источниками данных при испытаниях с прослушиванием, предназначенных для оценки небольших ухудшений качества в аудиосистемах (например, как описано в Рекомендации МСЭ-R BS.1116), должны быть оценщики, обладающие опытом определения таких небольших ухудшений. Чем более высокое качество звука достигается в испытуемых системах, тем важнее иметь опытных слушателей.

4.1 Критерии отбора оценщиков

Хотя метод испытания MUSHRA и не предназначен для оценки незначительных ухудшений, тем не менее рекомендуется привлекать опытных слушателей, с тем чтобы обеспечить надлежащее качество собираемых экспериментальных данных. Такие слушатели должны иметь опыт критического прослушивания звуковых материалов. Результат в этом случае будет более скорым и надежным, чем при использовании неопытных слушателей. Важно также отметить, что большинство неопытных слушателей, как правило, становятся чувствительнее к разнообразным артефактам после частого прослушивания. Опытного оценщика выбирают по умению выполнять соответствующие функции в ходе испытания с прослушиванием. Это умение должно быть качественно и количественно охарактеризовано путем испытания с повторным оцениванием по параметрам надежности и различительной способности:

– **различительная способность** – мера способности воспринимать различия между тестовыми элементами;

– **надежность** – мера близости повторных оценок одного и того же тестового элемента.

При окончательном анализе результатов следует использовать данные только от тех оценщиков, которые были классифицированы как *опытные*. Существует ряд методов характеризации оценщиков. Подробнее об этом см. в Отчете МСЭ-R BS.2300[1]. Эти методы предусматривают как минимум одну повторную оценку от каждого оценщика и позволяют качественно и количественно охарактеризовать

[1] Пример такого подхода – метод оценки квалификации (eGauge), описанный в Отчете МСЭ-R BS.2300-0. Отчет доступен в интернете по адресу http://www.itu.int/oth/R0A07000036.

уровень опытности оценщика в ходе одного эксперимента. Эти методы должны применяться либо на этапе первичного отбора оценщиков в рамках предварительного эксперимента, либо – что предпочтительнее – как на этапе первичного отбора, так и в ходе основного испытания. Предварительный эксперимент включает в себя серию экспериментов на репрезентативной подборке образцов, подлежащих оценке в основном эксперименте. Для оценки квалификации слушателя в предварительном эксперименте следует использовать релевантное подмножество тестовых входных сигналов, которое бы адекватно представляло все множество входных сигналов и артефактов, оцениваемых в ходе одного или нескольких основных экспериментов.

Графическое представление результатов этого анализа должно содержать информацию о надежности и различительной способности оценщиков.

4.1.1 Первичный отбор оценщиков

Группа оценщиков должна состоять из опытных слушателей, то есть тех, кто понимает описываемый метод субъективной оценки качества и был надлежащим образом обучен этому методу. Эти слушатели должны:

– обладать опытом критического прослушивания звуковых материалов;

– иметь нормальный слух (в этом отношении следует руководствоваться стандартом ISO 389).

Процедуру обучения следует использовать в качестве средства первичного отбора. При анализе результатов используются данные только от тех слушателей, которые классифицированы как *опытные оценщики* в ходе предварительного или основного экспериментов.

Повтор входных сигналов служит для оценки надежности слушателей.

Главный довод в пользу метода первичного отбора – повышение эффективности испытания с прослушиванием. Следует, однако, соотнести это соображение с риском чрезмерного ограничения релевантности результатов.

4.1.2 Последующее отсеивание оценщиков

Метод последующего отсеивания предусматривает исключение тех оценщиков, кто ставит очень высокую оценку существенно ухудшенному опорному сигналу, и тех, кто часто оценивает скрытый эталонный сигнал как существенно ухудшенный. Применяются следующие метрики:

– ответы оценщика следует исключить из общей совокупности, если он/она дает скрытому эталонному сигналу оценку ниже 90 баллов более чем для 15% тестовых элементов;

– ответы оценщика следует исключить из общей совокупности, если он/она дает опорному сигналу из среднего диапазона оценку выше 90 баллов более чем для 15% тестовых элементов. Если более 25% оценщиков дали опорному сигналу из среднего диапазона оценку выше 90 баллов, это может свидетельствовать о том, что тестовый элемент не претерпел значительного ухудшения в процессе обработки опорных сигналов. В этом случае не следует исключать оценщиков на основании оценок данного элемента.

При необходимости к этой начальной стадии можно приступить еще до того, как все оценщики завершат свою работу (это позволит испытательной лаборатории определить, имеется ли достаточное количество надежных оценщиков до завершения испытаний).

Может быть целесообразно изучить полученные данные для выявления в них ошибок, представляющих собой статистические выбросы, и подвергнуть такие выбросы дальнейшему анализу. Один из возможных методов – сравнение индивидуальных оценок с интерквартильным размахом всех оценок, выставленных конкретному условию испытаний j и последовательности звуковых элементов k.

Медиану $\hat{x}$ и квартили Q следует рассчитывать по следующим формулам:

$$\hat{x} \coloneqq Q_2(x_{jk}) = \text{median}(x)$$

$$\coloneqq \begin{cases} x_{jk\frac{n+1}{2}}, & n \text{ четное;} \\ \frac{1}{2}\left(x_{jk\frac{n}{2}} + x_{jk\frac{n}{2}+1}\right), & n \text{ нечетное,} \end{cases}$$

x упорядочено по возрастанию размера; и

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

$$Q_3(x_{jk}) = \begin{cases} \text{median}\left(x_{jk1}, \ldots, x_{jk\frac{n+1}{2}}\right), & n \text{ нечетное;} \\ \text{median}\left(x_{jk\frac{n}{2}+1}, \ldots, x_{jkn}\right), & n \text{ четное.} \end{cases}$$

Интерквартильный размах рассчитывается по следующей формуле: $IQR(x) \coloneqq Q_3(x) - Q_1(x)$.

В этом контексте выбросы принадлежат множеству $O(x_{jk})$:

$$O(x_{jk}) \coloneqq \left\{x_{jk} | x_{jk} > Q_3(x_{jk}) + 1{,}5 \cdot IQR(x_{jk})\right\} \cup \left\{x_{jk} | x_{jk} < Q_1(x_{jk}) - 1{,}5 \cdot IQR(x_{jk})\right\}.$$

Если оценка x, данная одним из участников конкретному входному сигналу и испытуемой системе, является элементом множества $O(x)$, следует выяснить причину такой оценки. В ходе исследования записи сеанса испытаний могут выявиться технические неполадки в оборудовании или ошибки, связанные с человеческим фактором. Возможно, опрос оценщика покажет, действительно ли выставленная оценка представляла его субъективное мнение. Если окажется, что выброс в данных обусловлен ошибкой, можно исключить этот выброс из набора данных перед окончательным анализом результатов с указанием причины исключения в протоколе испытания.

Применение метода последующего отсеивания может прояснить тенденции, наблюдаемые в результатах испытания. Вместе с тем, учитывая неодинаковую чувствительность оценщиков к различным артефактам, следует проявлять осторожность. Повысив численность группы оценщиков, можно уменьшить влияние оценок отдельных слушателей.

4.2 Численность группы оценщиков

Достаточная численность группы оценщиков может быть определена, если можно оценить дисперсию оценок различных участников группы и известна требуемая разрешающая способность эксперимента.

Опыт показывает, что, когда условия испытания с прослушиванием жестко контролируются как в техническом, так и в поведенческом отношении, данных не более чем от 20 участников зачастую бывает достаточно для того, чтобы сделать обоснованные выводы из испытания. Если есть возможность производить анализ по мере выполнения испытания оценщиками, то как только будет достигнут уровень статистической значимости, достаточный для получения обоснованных выводов, необходимость в обработке данных от большого количества оценщиков исчезает.

Если по той или иной причине жесткий контроль над экспериментом невозможен, то для достижения требуемой разрешающей способности может понадобиться большее количество оценщиков.

Численность группы оценщиков определяется не только требуемой разрешающей способностью. Результат эксперимента того типа, который рассматривается в настоящей Рекомендации, в принципе действителен только для той группы опытных слушателей, которая непосредственно участвовала в испытании. Таким образом, с повышением численности группы оценщиков можно утверждать, что результат будет действителен для более общей группы опытных слушателей и поэтому в некоторых случаях может считаться более убедительным. Необходимость в повышении численности группы оценщиков может также возникнуть в связи с вероятностью неодинаковой чувствительности оценщиков к различным артефактам.

5 Метод испытания

В методе испытания MUSHRA используют оригинальный необработанный материал программ во всей занимаемой им полосе частот, который играет роль эталонного сигнала (в том числе скрытого), и ряд обязательных скрытых опорных сигналов.

Могут также использоваться дополнительные скрытые опорные сигналы – желательно те, которые описаны в других Рекомендациях МСЭ-R, имеющих отношение к рассматриваемой здесь теме. Поскольку свойства опорных сигналов могут в значительной степени влиять на результат испытания, при проектировании нестандартного опорного сигнала следует учитывать требования к оптимальному поведению опорных сигналов, изложенные в Приложении 5. В протоколе испытания должен быть подробно описан характер используемых в нем нестандартных опорных сигналов.

5.1 Описание тестовых сигналов

Рекомендованная максимальная длительность последовательностей – 10 с, желательно не более 12 с. Это нужно для того, чтобы не утомлять слушателей, повысить надежность и стабильность их ответов, а также сократить общую продолжительность испытания с прослушиванием. Такая длительность также необходима для обеспечения единообразия контента на всем протяжении сигнала, вследствие чего ответы слушателей также должны стать более однородными. Кроме того, меньшая длительность позволит слушателям сравнивать более длительные в относительном выражении непрерывные отрезки тестовых сигналов.

Если сигналы слишком длительные, ответы слушателей определяются эффектами первичности и недавности тестовых сигналов или изолированными закольцованными фрагментами, спектральные и временные характеристики которых могут сильно различаться на протяжении тестового сигнала. Длительность тестовых сигналов сокращают с целью уменьшить этот разброс. Однако в некоторых условиях это ограничение может оказаться неуместным. Пример – испытание с длинной медленно движущейся траекторией звука. В тех частных случаях, когда обнаруживается потребность в более длительном входном сигнале, необходимо документально зафиксировать обоснование такой потребности в окончательном протоколе испытания.

Набор обрабатываемых сигналов состоит из всех тестовых сигналов и как минимум двух дополнительных опорных сигналов. Стандартный опорный сигнал представляет собой исходный сигнал, пропущенный через фильтр нижних частот с частотой среза 3,5 кГц, а опорный сигнал среднего качества – то же, но с частотой среза 7 кГц.

Полосы частот тестовых сигналов соответствуют Рекомендациям для каналов управления (3,5 кГц), используемых в радиовещании для целей контроля и координации, комментаторских каналов (7 кГц) и заказываемых каналов (10 кГц) – Рекомендации МСЭ-T G.711, G.712, G.722 и J.21 соответственно.

Фильтр нижних частот с частотой среза 3,5 кГц должен иметь следующие характеристики:

– f_c = 3,5 кГц;

– неравномерность АЧХ в полосе пропускания – не более ±0,1 дБ;

– затухание на частоте 4 кГц – не менее 25 дБ;

– затухание на частоте 4,5 кГц – не менее 50 дБ.

Назначение дополнительных опорных сигналов – показать, как звучание испытуемой системы соотносится с хорошо известными уровнями качества. Эти сигналы не следует использовать для перенормировки результатов между различными испытаниями.

5.2 Этап обучения

Для получения надежных результатов в обязательном порядке следует проводить обучение оценщиков, для чего перед испытанием организуются специальные учебные сеансы. Установлено, что такое обучение играет важную роль в обеспечении надежности экспериментальных результатов. Как минимум в ходе обучения участник должен быть ознакомлен с возможными ухудшениями во всем их качественном и количественном разнообразии, а также со всеми тестовыми сигналами, которые будут представлены во время испытания. Этого можно достичь разными способами,

например с помощью простой магнитофонной системы или интерактивной системы с компьютерным управлением. Инструкции даны в Приложении 1. Кроме того, обучение должно обеспечить знакомство оценщиков с установкой для субъективных испытаний (например, с соответствующим программным обеспечением).

5.3 Представление входных сигналов

MUSHRA – двойной слепой метод с множеством входных сигналов, в котором используются скрытый эталонный сигнал и скрытые опорные сигналы, тогда как в Рекомендации МСЭ-R BS.1116 применяется "двойной слепой метод с тремя входными сигналами и скрытым эталонным сигналом". Подход, принятый в методе MUSHRA, представляется более подходящим для оценки средних и значительных ухудшений [4].

В испытании с небольшими ухудшениями сложность для участника состоит в обнаружении артефактов, которые могут присутствовать в сигнале. В этой ситуации необходимо, чтобы в ходе испытания воспроизводился скрытый эталонный сигнал, это позволит экспериментатору оценить умение оценщика обнаруживать артефакты. И наоборот, в испытании со средними и значительными ухудшениями участник не испытывает затруднений с выявлением артефактов, поэтому скрытый эталонный сигнал не нужен. Трудности скорее возникают тогда, когда участнику приходится оценивать относительные уровни неудобств, доставляемых различными артефактами. Здесь участник должен ранжировать разные типы артефактов согласно своим предпочтениям.

В связи с применением высококачественного эталонного сигнала возникает интересная проблема. Поскольку новая методика предназначается для оценки средних и значительных ухудшений, воспринимаемая разница между эталонным сигналом и тестовыми элементами предполагается относительно большой. При этом воспринимаемая разница между тестовыми элементами, принадлежащими к различным системам, может быть совсем невелика. В итоге если используется метод с множественными испытаниями (например, как в Рекомендации МСЭ-R BS.1116), точное распознавание разнообразных ухудшенных сигналов может представлять большую трудность для оценщиков. Например, в испытании с непосредственным попарным сравнением оценщики могут сойтись в том, что система A лучше системы B. Но когда звучание каждой системы сравнивается только с эталонным сигналом (то есть системы A и B не сравниваются непосредственно друг с другом), информация о различиях между двумя системами может быть утрачена.

Для преодоления этой трудности в методе MUSHRA для участника предусмотрена возможность произвольно переключаться между эталонным сигналом и любой из испытуемых систем (обычно при помощи проигрывателя с компьютерным управлением, хотя может быть также использовано несколько проигрывателей компакт-дисков или магнитофонов). Участник принимает участие в последовательности испытаний. В ходе каждого испытания вниманию участника представляются эталонный сигнал, опорные сигналы низкого и среднего качества, а также версии тестового сигнала, обработанные испытуемыми системами. Например, если испытываются 8 аудиосистем, участнику предоставляется возможность практически мгновенно переключаться между 11 тестовыми сигналами и явным эталонным сигналом (1 эталонный сигнал + 8 сигналов испытуемых систем + 1 скрытый эталонный сигнал + 1 скрытый опорный сигнал низкого качества + 1 скрытый опорный сигнал среднего качества).

Поскольку участник может непосредственно сравнивать ухудшенные сигналы, данный метод обладает преимуществами испытания с полным попарным сравнением в том смысле, что участнику легче выявлять различия между ухудшенными сигналами и выставлять им соответствующие оценки. Данная особенность позволяет оценивать системы с высокой разрешающей способностью. Важно, однако, отметить, что при выставлении оценки конкретной системе оценщики будут исходить из сравнения звучания этой системы с эталонным сигналом, а также с другими сигналами в ходе каждого испытания.

Рекомендуется в каждое испытание включать не более 12 сигналов (например, сигналы 9 испытуемых систем, 1 скрытый опорный сигнал низкого качества, 1 скрытый опорный сигнал среднего качества и 1 скрытый эталонный сигнал).

В редких случаях, когда стоит задача сравнить большое количество сигналов, может потребоваться блочный план эксперимента, который должен быть подробно описан в протоколе.

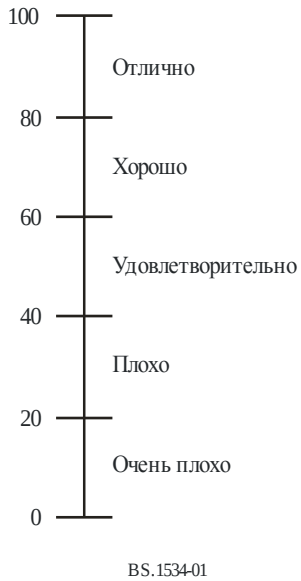
При выполнении испытаний согласно Рекомендации МСЭ-R BS.1116 оценщики обычно начинают испытание с выявления артефактов, а затем переходят к оцениванию. Опыт проведения испытаний по методу MUSHRA показывает, что оценщики, как правило, начинают сеанс с грубой оценки качества. Затем следует сортировка или ранжирование. После этого участники выставляют оценки. Поскольку ранжирование производится напрямую, результаты в случае промежуточного качества звука получаются более однородными и надежными, чем если бы использовался метод из Рекомендации МСЭ-R BS.1116. Минимальная длительность закольцованного фрагмента составляет 500 мс, и ко всем закольцованным фрагментам следует применить эффект плавного нарастания в начале и затухания в конце на участке длительностью 5 мс с огибающей типа приподнятого косинуса. При любых переключениях контента между испытуемыми системами уровень сигнала также должен плавно нарастать в начале и затухать в конце по огибающей типа приподнятого косинуса на участке длительностью 5 мс. Ни при каких обстоятельствах в ходе испытаний не следует использовать плавный переход одного сигнала в другой при переключении между испытуемыми системами. Цель этих модификаций – уменьшить число изменений спектральной окраски при сравнениях с резкими переходами для более качественной идентификации и оценки испытуемых сигналов.

5.4 Процесс оценивания

Перед оценщиками ставится задача оценить входные сигналы по непрерывной шкале качества (CQS). Шкала CQS состоит из идентичных вертикальных графических шкал (длиной обычно 10 см или более), которые разделены на 5 равных отрезков, обозначенных определениями (рисунок 1).

Эта шкала также используется для оценки качества изображений (Рекомендация МСЭ-R BT.500 "Методика субъективной оценки качества телевизионных изображений").

РИСУНОК 1



BS.1534-01

Слушатель фиксирует свою оценку качества в подходящей форме, например с помощью бегунков на электронном дисплее (рисунок 2) или карандашом на бумажной шкале. С помощью установки, сходной с показанной на рисунке 2, следует ограничить действия участника так, чтобы он мог изменять оценку только того элемента, который он прослушивает в данный момент. Некоторые руководящие указания по проектированию интерфейсов можно найти в Приложении 2. Оценщика просят оценить качество всех входных сигналов по шкале CQS из пяти отрезков.

РИСУНОК 2

Пример изображения на дисплее компьютера в ходе испытания по методу MUSHRA

BS.1534-02

Преимущество метода MUSHRA в сравнении Рекомендацией МСЭ-R BS.1116 – одновременное воспроизведение множества входных сигналов, что позволяет участнику непосредственно сравнивать любые из них. Испытание по методу MUSHRA может быть проведено за существенно меньшее время, чем по методу из Рекомендации МСЭ-R BS.1116.

5.5 Запись сеансов испытаний

На случай обнаружения каких-либо аномалий при обработке выставленных оценок полезно иметь запись событий, следствием которых стали эти оценки. Относительно простой способ добиться этого – видео- и аудиозапись всего испытания. В случае когда набор результатов содержит аномальную оценку, можно проанализировать запись и попытаться установить, в чем была причина – в технических неполадках или в человеческом факторе.

6 Атрибуты

Ниже перечислены атрибуты, специфичные для оценки монофонических, стереофонических и многоканальных систем. Предпочтительно в каждом случае оценивать атрибут "базовое качество звука". Экспериментаторы по своему выбору могут определять и оценивать другие атрибуты.

В ходе испытания следует оценивать только один атрибут. Когда оценщиков просят оценить более одного атрибута в каждом испытании, попытки ответить на множество вопросов об одном входном сигнале могут привести к перегрузке внимания, замешательству или тому и другому одновременно. Результатом могут стать ненадежные оценки по всем заданным вопросам. Если стоит задача дать независимую оценку множеству характеристик звучания, рекомендуется вначале оценить базовое качество звука.

6.1 Монофоническая система

Базовое качество звука – этот единый глобальный атрибут служит для оценки всех обнаруженных различий между эталоном и объектом испытаний.

6.2 Стереофоническая система

Базовое качество звука – этот единый глобальный атрибут служит для оценки всех обнаруженных различий между эталоном и объектом испытаний.

Ниже приведен дополнительный атрибут, который может также представлять интерес.

Качество стереофонического образа – этот атрибут характеризует различия между эталоном и объектом испытаний по параметрам местоположения звукового образа и ощущений глубины и реальности звукового события. Хотя в некоторых исследованиях было показано, что качество стереофонического образа может ухудшаться, имеющихся данных недостаточно, чтобы установить обоснованность оценки качества стереофонического образа отдельно от базового качества звука.

ПРИМЕЧАНИЕ 1. – До 1993 года в большинстве исследований по субъективной оценке небольших ухудшений качества в стереофонических системах оценивался только атрибут "базовое качество звука". Соответственно в этих исследованиях атрибут "качество стереофонического образа" явным или неявным образом включался в глобальный атрибут "базовое качество звука".

6.3 Многоканальная система

Базовое качество звука – этот единый глобальный атрибут служит для оценки всех обнаруженных различий между эталоном и объектом испытаний.

Ниже приведены дополнительные атрибуты, которые могут также представлять интерес.

Качество фронтального образа – этот атрибут связан с локализацией фронтальных источников звука. Он включает в себя качество стереофонического образа и потери четкости.

Ощущение качества объемного звука – этот атрибут связан с ощущениями пространства, объемностью или специальными эффектами направленности в объемном звучании.

7 Тестовый материал

Для выявления различий между испытуемыми системами следует использовать критичный материал, представляющий типичную для рассматриваемого применения программу радиовещания. Критичный материал – это такой материал, который находится на пределе возможностей испытуемой системы. Не существует какого-либо универсально пригодного программного материала, который можно было бы использовать для оценки всех систем в любых условиях. Соответственно, необходимо специально изыскивать критичный программный материал для каждой системы, подлежащей испытанию в каждом эксперименте. Как правило, поиск подходящего материала – трудоемкая работа, но если не найти по-настоящему критичный материал для каждой системы, то в ходе экспериментов не удастся выявить различия между системами и сделать какие-либо выводы будет нельзя. Следует поручить небольшой группе слушателей-экспертов отобрать тестовые элементы из более обширного подходящего набора. Процесс отбора должен охватывать все испытуемые системы и должен быть задокументирован в протоколе испытания.

Необходимо эмпирически и статистически показать, что невыявление различий между системами не может быть обусловлено нечувствительностью самого эксперимента, связанной с неудачным выбором аудиоматериала или другими недочетами в эксперименте. В противном случае такой нулевой результат не может быть признан действительным.

В поисках критичного материала допустимо рассматривать любой входной сигнал, потенциально пригодный для использования в радиовещательной программе. Синтетические сигналы, специально спроектированные как нерасчетные для конкретной системы, использовать для этих целей не следует. Художественное или интеллектуальное содержимое последовательности программ не должно быть чрезмерно привлекательным или, наоборот, неприятным либо утомительным, с тем

чтобы не отвлекать участника от выявления ухудшений. Следует учитывать ожидаемую частоту появления каждого типа программных материалов в реальных радиовещательных программах. Вместе с тем следует понимать, что со временем характер материалов может измениться со сменой музыкальных стилей и предпочтений.

При выборе программного материала важно точно определить атрибуты, подлежащие оцениванию. Отбор материала следует поручать группе опытных оценщиков, обладающих базовыми знаниями об ожидаемых ухудшениях. Изначально им предлагается очень широкое разнообразие материала, которое можно дополнительно расширить за счет специально сделанных записей.

Для подготовки к формальному субъективному испытанию необходимо перед записью каждого фрагмента на тестовый носитель поручить группе опытных оценщиков отрегулировать громкость фрагмента по субъективным ощущениям. Это позволит затем в ходе испытания воспроизводить все элементы программы с тестового носителя при фиксированной установке усиления.

Для всех тестовых последовательностей группа опытных оценщиков собирается для достижения консенсуса по поводу относительных уровней отдельных тестовых фрагментов. Кроме того, эксперты должны прийти к единому мнению об абсолютном уровне звукового давления при воспроизведении всей последовательности в целом относительно уровня выравнивания.

В начале каждой записи может присутствовать тональная посылка (например, с частотой 1 кГц, длительностью 300 мс и уровнем –18 дБ полной шкалы) на уровне сигнала выравнивания, позволяющая установить выходной уровень выравнивания равным входному уровню выравнивания, требуемому в канале воспроизведения, согласно Рекомендации EPC R.68 (см. Рекомендацию МСЭ-R BS.1116, п. 8.4.1). Тональная посылка предназначена только для выравнивания и не должна воспроизводиться в ходе испытания. Сигнал звуковой программы следует отрегулировать так, чтобы амплитуды его пиков лишь в редких случаях превышали пиковую амплитуду максимально допустимого сигнала, определенного в Рекомендации МСЭ-R BS.645 (синусоидальный сигнал с уровнем на 9 дБ выше уровня выравнивания).

Практически целесообразное количество фрагментов в испытании может различаться, но должно быть одинаковым для всех испытуемых систем. Разумная оценка – количество испытуемых систем, умноженное на 1,5, но не менее 5 фрагментов. Ввиду сложности задачи испытуемые системы должны быть доступны экспериментатору. Успешный отбор возможен только при правильно составленном графике. Кроме того, ввиду использования переменной скорости передачи данных в аудиокодеках рекомендуется кодировать более длинные последовательности и задействовать часть каждой последовательности в испытании с прослушиванием.

Характеристики многоканальной системы в условиях двухканального воспроизведения определяются по эталонному понижающему микшированию. Использование фиксированного понижающего микширования в некоторых обстоятельствах может с некоторых точек зрения ограничивать возможности, но это, без сомнения, наиболее разумный вариант на долгосрочную перспективу для радиовещательных организаций. Уравнениями для эталонного сведения (см. Рекомендацию МСЭ-R BS.775) являются:

$$L_0 = 1{,}00L + 0{,}71C + 0{,}71L_S;$$

$$R_0 = 1{,}00R + 0{,}71C + 0{,}71R_S.$$

Предварительный отбор тестовых фрагментов для критической оценки характеристик эталонного двухканального понижающего микширования следует производить на основе воспроизведения программного материала, смикшированного в два канала.

8 Условия прослушивания

Методы субъективной оценки малых искажений в аудиосистемах, включая многоканальные звуковые системы, определены в Рекомендации МСЭ-R BS.1116. Условия прослушивания для оценки аудиосистем с промежуточным качеством звука следует устанавливать в соответствии с разделами 7 и 8 Рекомендации МСЭ-R BS.1116.

Испытание может проводиться с использованием либо наушников, либо громкоговорителей. Не допускается одновременное использование того и другого в одном сеансе испытания: все оценщики должны пользоваться преобразователем одного и того же типа.

Для измерительного сигнала со среднеквадратичным напряжением, равным "уровню сигнала выравнивания" (0 дБн0з согласно Рекомендации МСЭ-R BS.645; –18 дБ ниже уровня срезания пиков уровня цифровой записи согласно Рекомендации ЕРС R.68), поданного поочередно на вход каждого канала воспроизведения (то есть усилителя мощности с соответствующим громкоговорителем), коэффициент усиления усилителя устанавливается с таким расчетом, чтобы эталонный уровень звукового давления (взвешенный по МЭК-А для случая медленного изменения) составлял

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

где n – общее число каналов воспроизведения в испытательной установке.

Допускается индивидуальная регулировка уровня прослушивания участником в пределах сеанса. Ее следует ограничить диапазоном ±4 дБ относительно эталонного уровня, определенного в Рекомендации МСЭ-R BS.1116. Группе, производящей отбор, следует обеспечить такой баланс между тестовыми элементами в рамках одного испытания, чтобы оценщикам в обычных обстоятельствах не приходилось индивидуально регулировать уровень каждого элемента.

Регулировка уровня по ходу воспроизведения конкретного элемента не допускается.

9 Статистический анализ

Оценки по каждому из условий испытаний линейно преобразуются в нормализованные оценки в баллах по шкале от 0 до 100, где 0 соответствует нижней границе шкалы (плохое качество). Затем рассчитываются абсолютные оценки, как описано ниже.

Может выполняться параметрический либо непараметрический статистический анализ исходя из выполнения статистических предположений (см. п. 9.3.3). Руководящие указания по параметрическому статистическому анализу см. в Приложении 4.

9.1 Визуализация и исследовательский анализ данных

Статистический анализ должен всегда начинаться с визуализации исходных данных. Для этой цели могут использоваться гистограммы с кривой аппроксимации для нормального распределения, коробчатые диаграммы или диаграммы квантиль-квантиль (Q-Q).

Коробчатая диаграмма показывает наличие выбросов и их влияние на сводные описательные показатели. Этот вид визуализации следует использовать для определения разброса и отклонений отдельных оценок от медианы по всем оценщикам. Визуализацию с помощью гистограммы следует производить для выявления многомодальности исходного распределения данных. Если визуализация четко демонстрирует наличие многомодального распределения, экспериментатору рекомендуется проанализировать его отдельно.

Оценить степень многомодальности b можно по следующей формуле:

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

где:

n: размер выборки;

g: коэффициент асимметрии конечной выборки;

k: эксцесс результатов испытания с прослушиванием.

Этот коэффициент принимает значения от 0 до 1. Высокие его значения (> 5/9) можно интерпретировать как признак многомодальности.

Исходя из вида этих диаграмм, значения коэффициента b и предположений об исходной совокупности, лежащей в основе наблюдаемой выборки, следует решить, можно ли предполагать, что наблюдается нормальное распределение. Если кривая аппроксимации явно асимметрична, на гистограмме присутствует множество выбросов или диаграмма квантиль-квантиль существенно отличается от прямой линии, не следует считать выборку нормально распределенной. Вычисление медианы нормализованных оценок от всех слушателей, успешно прошедших последующее отсеивание, дает медианную субъективную оценку.

Медиана рассчитывается по формуле

$$\hat{x} = \text{median}(x) = \begin{cases} x_{\frac{n+1}{2}}, & n \text{ нечетное;} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right), & n \text{ четное значение } x \text{ упорядочено по возрастанию размера.} \end{cases}$$

Первый шаг анализа – вычисление медианной оценки $\overline{\eta}_{jk}$ для каждого из представлений. Следовательно, η_{ijk} есть медианная оценка наблюдателя i для данного условия испытания j и звуковой последовательности k, а $\hat{\eta}$ – медиана по всей выборке (все наблюдатели, все условия, все звуковые последовательности).

Аналогичным образом можно рассчитать общие медианные оценки $\overline{\eta}_j$ и $\overline{\eta}_k$ для каждого из условий испытаний и каждой тестовой последовательности.

Хотя некоторые методы анализа, например ANOVA (см. п. 9.3), требуют расчета средних значений, характеризовать положение центра распределения можно также с помощью медианы. Медиана – робастная характеристика положения центра распределения, оптимальная при малом размере выборки, ненормальном распределении или наличии значимых выбросов. Вероятно, во многих экспериментальных сценариях эти соображения не столь существенны. Но поскольку одним из главных преимуществ стандартизированных испытаний являются сравнение и интерпретация оценок, данных разными пользователями в разных лабораториях, полезно идентифицировать методы анализа, которые наиболее робастны и наименее чувствительны к факторам, способным повлиять на действительность результатов или возможность их сравнения.

В связи с этим можно применять непараметрические статистические методы. При непараметрическом анализе данных средние значения и доверительные интервалы для уровня значимости 95% следует рассчитывать доступными методами, например по распространенному алгоритму бутстрапирования.

Меры отклонений от медианы можно рассчитывать, используя среднее абсолютное отклонение:

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

В качестве характеристики разброса распределения вокруг медианы рекомендуется использовать интерквартильный размах (IQR). Это разность между третьим и первым квартилями: $IQR = Q_3 - Q_1$. Эти формулы приведены в п. 4.1.2. Если результаты распределены нормально, IQR равен удвоенному среднему абсолютному отклонению.

Определять статистическую значимость рекомендуется на уровне значимости 95%. Непараметрические критерии рандомизации – робастные характеристики статистической значимости. В отличие от параметрических статистических величин они не содержат предположений об исходном распределении данных и менее чувствительны ко многим факторам, связанным с использованием выборок меньшего размера.

Робастный непараметрический критерий рандомизации (критерий перестановок) позволяет определить вероятность того, что наблюдаемое различие между двумя условиями испытания имело бы место при истинно случайных данных, как предполагается в нулевой гипотезе. Вероятность, вычисленная по этому критерию, – реальная характеристика, определенная по фактическому распределению данных, а не гипотетическая характеристика, предполагающая определенную форму

исходного распределения [5]. Такого рода проверки требуют применения общепринятых методов перевыборки, таких как бутстрапирование и численный метод Монте-Карло, что сегодня стало реальностью благодаря высокому быстродействию современных компьютеров [6]. Более подробное описание этого метода проверки статистических гипотез дано в Приложении 3.

9.2 Анализ мощности

Анализ мощности в *априорном* варианте может быть полезным средством для оценки необходимого размера выборки для испытаний с прослушиванием, а в *апостериорном* – для оценки мощности критерия, связанной с вероятностью совершения ошибок второго рода. *Априорный* анализ дает необходимый размер выборки для эксперимента, если известны величина эффекта $d = \frac{\bar{x}}{s}$, уровень значимости α и мощность критерия $1 - \beta$.

Апостериорный анализ, наоборот, позволяет определить мощность критерия $1 - \beta$ или вероятность ошибки второго рода β при использовании этого критерия, зная величину эффекта $d = \frac{\bar{x}}{s}$, уровень значимости α и размер выборки N. Вероятность ошибки второго рода β – это вероятность того, что эффект d присутствует в выборке, но не будет признан значимым по применяемому критерию. Если, например, согласно некоторому критерию система не влияет на качество, $1 - \beta$ описывает вероятность того, что ухудшение действительно не имело места[2].

9.3 Применение и использование метода ANOVA

9.3.1 Введение

В данном подразделе рассматриваются требования к параметрической статистической обработке методом дисперсионного анализа (ANOVA). Ввиду своей робастности и статистической мощности[3] модель ANOVA хорошо подходит для обработки данных, собранных с использованием методов, описанных в Рекомендации МСЭ-R BS.1534 (см. [7], [8], [12], [13]). Поскольку F-статистика метода ANOVA робастна как к ненормальности распределения, так и гетерогенности дисперсии, проверке подлежат главным образом предположения о характере ошибок (в том числе остаточных).

Подробнее об общих предположениях, связанных с параметрическим статистическим анализом, см. в Приложении 4.

9.3.2 Определение модели

Настоятельно рекомендуется на этапе планирования эксперимента (см. раздел 3) всесторонне определить модель в терминах независимых переменных (например, ОБРАЗЕЦ, СИСТЕМА, УСЛОВИЕ и т. д.) и зависимых переменных (например, "базовое качество звука", "усилия при прослушивании" и т. д.). В модели следует также определить уровни каждой независимой переменной.

Важно, чтобы в определяемую аналитическую модель (например, дисперсионного анализа ANOVA или повторяющихся измерений ANOVA) были включены все значимые переменные. Упущение значимых переменных, например двух- и трехсторонних взаимодействий независимых факторов, может привести к некорректному определению модели, а это в свою очередь к плохо объяснимой дисперсии (R^2) и неправильной интерпретации результатов анализа данных.

9.3.3 Контрольный список для параметрического статистического анализа

Приведенный здесь контрольный список призван служить кратким руководством по первичному анализу данных, проверке основных предположений (параметрического и непараметрического характера), а также проведению основных этапов параметрического статистического анализа.

---

[2] Существует множество средств для автоматического анализа мощности для известных распределений, например G*Power [16]. Оценивать мощность критериев для неизвестных распределений сложнее.

[3] В общем случае рекомендуется выбирать наиболее мощный метод статистического анализа, допустимый исходя из характера имеющихся данных [9], [10].

Основной упор в контрольном списке делается на требованиях к дисперсионному анализу как надлежащему методу анализа экспериментальных данных, полученных согласно методикам, описанным в Рекомендации МСЭ-R BS.1534. За исчерпывающими указаниями читателю рекомендуется обращаться к учебникам по статистике (например, [8], [11], [9]).

- Исследовательская статистика[4]
  - Проверить правильность структуры данных и ее соответствие ожиданиям;
  - проверить, нет ли недостающих данных;
  - исследовать нормальность распределений данных;
  - рассмотреть другие потенциальные распределения данных (унимодальные, бимодальные, асимметричные и т. д.).
- Одномерность
  - Проверить использование всеми оценщиками одной и той же шкалы[5];
  - проверить, одномерны ли данные по своему характеру;
  - метод главных компонент, диаграммы Такера-1 или коэффициент альфа Кронбаха.
- Независимость наблюдений
- Обычно определяется методологией эксперимента и не может быть легко проверена статистическими методами. Следует обеспечить независимый сбор данных, например, применив двойные слепые методы и исключив взаимовлияние оценщиков.
- Однородность дисперсии[6]
  - Проверить предположение о том, что каждая независимая переменная демонстрирует сходную дисперсию:
    - провести визуальную проверку, построив рядом коробчатые диаграммы для каждого уровня независимых переменных; согласно практическому правилу, гетерогенность обусловливает не более чем четырехкратную разницу;
    - для оценки однородности дисперсии можно применять критерий Брауна-Форсайта или статистику Левена.
- Нормальное распределение остаточных ошибок
  - Проверить нормальность распределения остаточных ошибок:
    - D-критерий Колмогорова-Смирнова, K-S-критерий Лиллефорса или критерий Левена;
    - визуально оценить нормальность распределения можно также по графику нормального распределения вероятности (иногда называется P-P-графиком) или диаграмме квантиль-квантиль (часто называется Q-Q-диаграммой).
- Выявление выбросов
  - Следует выявить и, возможно, исключить выбросы в случаях, когда это оправданно. Руководящие указания на этот счет приведены в п. 4.1.2.
- Анализ
  - Дисперсионный анализ (ANOVA) – общая линейная модель или модель повторяющихся измерений ANOVA:
    - применить подходящую модель ANOVA, например общую линейную модель (GLM) или модель повторяющихся измерений; подробнее см. в Приложении 4;
    - задать модель в соответствии с планом эксперимента:

---

[4] Применимо равным образом к параметрическому и непараметрическому статистическому анализу.

[5] Многомерность наблюдалась в случаях, когда разные подгруппы общей совокупности расходились во мнениях относительно оценки конкретных артефактов.

[6] Требуется для применения метода ANOVA, но не rmANOVA (см. Приложение 4).

- по возможности учесть двух- и трехсторонние взаимодействия;

– проанализировать данные с использованием модели и результатов:

- проанализировать объяснимую дисперсию (R^2) модели, используемой для описания зависимой переменной;
- проанализировать распределение остаточной ошибки;
- проанализировать значимые и незначимые факторы;

– модель можно определять в несколько итераций для исключения выбросов и незначимых факторов.

– Апостериорные проверки

– С помощью апостериорных критериев установить значимость различий между средними значениями в случаях, когда зависимый фактор (или взаимодействие факторов) значимы в модели ANOVA;

– существует ряд апостериорных критериев с разными уровнями различения, например критерий наименьшей значимой разности Фишера (LSD), критерий достоверно значимой разности Тьюки (HSD) и т. д.;

– рекомендуется указывать в протоколе величины эффектов вместе с уровнями значимости.

– Выводы

– После выполнения анализа подвести его итоги, изобразив графически средние значения и соответствующие доверительные интервалы по уровню значимости 95% для исходных или смоделированных по ANOVA данных (иногда называемые оцененными пределами средних).

– В случаях, когда взаимодействия факторов (например, двух- или трехсторонние) признаются значимыми, следует изобразить их графически, чтобы дать всестороннее представление о данных. Если в таких случаях изобразить только основные эффекты, представление получится неполным, так как будет нивелирован эффект взаимодействия.

Дальнейшие указания по использованию моделей ANOVA можно найти в Приложении 4, а также в популярной статистической и прикладной литературе, например, [11], [13], [15].

10 Протокол и представление результатов испытаний

10.1 Общие замечания

Представлять результаты следует в удобном для восприятия виде, с тем чтобы любой читатель, будь то неопытный или высококвалифицированный, смог получить нужную информацию. Прежде всего любого читателя интересует общий итог эксперимента, предпочтительно представленный в графической форме. Такое представление может сопровождаться более подробной информацией количественного характера, хотя развернутый численный анализ следует выносить в приложения.

10.2 Содержание протокола испытаний

В протоколе испытаний должны быть как можно более ясно изложены обоснования исследования, примененные методы и сделанные выводы. Изложение должно быть достаточно подробным, с тем чтобы квалифицированный специалист мог в принципе воспроизвести исследование для эмпирической проверки его результатов. Вместе с тем нет нужды включать в протокол каждый отдельный результат. Необходимо, чтобы информированный читатель был способен понять и подвергнуть критическому анализу основные аспекты испытания, в частности его мотивировку, методы планирования и выполнения эксперимента, методы анализа и выводы.

Особое внимание необходимо уделить следующему:

– графическое представление результатов;

– графическое представление отбора *опытных оценщиков* и требований к ним;

– определение плана эксперимента;

– требования к тестовому материалу и его отбор;

– общие сведения о системе, использованной для обработки тестового материала;

– подробная схема проведения испытания;

– детальные физические характеристики среды и оборудования для прослушивания, в том числе размеры и акустические характеристики помещения, типы и местоположения акустических преобразователей, характеристики электрооборудования (см. Примечание 1);

– план эксперимента, обучение, инструкции, последовательности экспериментов, методики испытаний, генерация данных;

– обработка данных, в том числе подробная описательная статистика и аналитическая выводная статистика;

– факт использования опорных сигналов в ходе испытания;

– методы последующего отсева, использованные при анализе результатов, включая методы исключения выбросов и неподготовленных слушателей;

– Рекомендация, в соответствии с которой выполнялись испытания (МСЭ-R BS.1534 или МСЭ-R BS.1534-1); ее следует четко указать в документе с описанием использованных опорных условий;

– корректный код для определения и генерации опорных сигналов, который бы позволял новому пользователю получить любой примененный в ходе испытаний опорный сигнал, не описанный явно в настоящей Рекомендации МСЭ-R BS.1534-2;

– подробно изложенные основания всех сделанных выводов.

ПРИМЕЧАНИЕ 1. – Поскольку есть отдельные свидетельства, что условия прослушивания, например вид акустического преобразователя (громкоговоритель или наушники), могут влиять на субъективную оценку, желательно, чтобы экспериментаторы явно указывали эти условия и тип звуковоспроизводящего оборудования, использованного в ходе экспериментов. Если есть намерение провести комбинированный статистический анализ преобразователей разных типов, необходимо убедиться в том, что такое объединение результатов возможно.

10.3 Представление результатов

Для каждого экспериментального параметра должны быть приведены медиана и интерквартильный размах (IQR) статистического распределения оценок.

Результаты должны сопровождаться следующей информацией:

– описание тестовых материалов;

– количество оценщиков;

– графическое представление результатов; коробчатые диаграммы с указанием IQR, а также средние значения и доверительные интервалы по уровню значимости 95%; следует сообщить о существенных различиях между испытуемыми системами, а также указать примененный метод статистического анализа.

За коробчатой диаграммой может следовать представление результатов в других формах, допускаемых характером данных, например в виде средних значений и доверительных интервалов.

10.4 Абсолютные оценки

Представление абсолютных средних оценок испытуемых систем, скрытого эталонного сигнала и опорных сигналов адекватно характеризует общие результаты испытания. Следует однако иметь в виду, что такое представление не содержит подробной информации о проведенном статистическом анализе. Следовательно, наблюдения не являются независимыми, и статистический анализ одних лишь абсолютных оценок без учета исходной совокупности, лежащей в основе наблюдаемой выборки, не дает осмысленной информации. Кроме того, следует указать примененные статистические методы, как предложено в разделе 9.

10.5 Уровень значимости и доверительный интервал

В протоколе испытания следует привести информацию о статистической природе всех субъективных данных. Следует указать уровни значимости, а также другие сведения о статистических методах и результатах их применения, это облегчит читателю понимание. Такие сведения могут включать, например, доверительные интервалы или планки погрешностей на графиках.

Разумеется, не существует какого-то "правильного" уровня значимости, но традиционно выбирается значение 0,05. Теоретически возможно использование односторонних или двусторонних критериев в зависимости от проверяемой статистической гипотезы.

Справочные документы

[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] МСЭ-Т. Практические процедуры для проведения субъективных испытаний. Международный союз электросвязи. 2011 год.

[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.

Прилагаемый документ 1
к Приложению 1
(нормативное)

Инструкция оценщикам

Ниже приведен типовой образец инструкций по проведению испытаний, которые следует выдавать в письменном виде или зачитывать оценщикам.

1 Этап ознакомления (обучения)

Первый этап в испытаниях с прослушиванием – знакомство с процессом испытаний. Он называется этапом обучения и предшествует этапу формальной оценки.

Этап обучения нужен для того, чтобы помочь вам как оценщику:

– ознакомиться со всеми звуковыми фрагментами, включенными в испытание, и соответствующими диапазонами уровней качества (**стадия A**); а также

– научиться пользоваться испытательным оборудованием и шкалой оценки (**стадия B**).

На стадии A этапа обучения вы сможете прослушать все отобранные для испытаний звуковые фрагменты, с тем чтобы оценить весь диапазон возможных характеристик звука. Представленные вашему вниманию звуковые фрагменты, которые будут более или менее критичными в зависимости от скорости передачи данных и других "условий". На рисунке 3 показан пользовательский интерфейс. Нажимая различные кнопки, вы сможете прослушивать разные звуковые фрагменты, в том числе эталонные. Так вы научитесь оценивать диапазон уровней качества для разных программных элементов. Фрагменты сгруппированы на основании общих условий. В данном случае выделены три такие группы. В каждую группу включено по четыре обработанных сигнала.

На стадии B этапа обучения вы научитесь пользоваться оборудованием для воспроизведения и оценки качества звуковых фрагментов, с которым будете иметь дело в ходе испытания.

На этапе обучения вы научитесь давать собственную оценку ухудшениям качества звука по предлагаемой шкале. Не следует обсуждать вашу личную интерпретацию шкалы с другими оценщиками на этапе обучения. Вместе с тем приветствуется разъяснение артефактов другим оценщикам.

Оценки, выставленные на этапе обучения, не будут учитываться в ходе реальных испытаний.

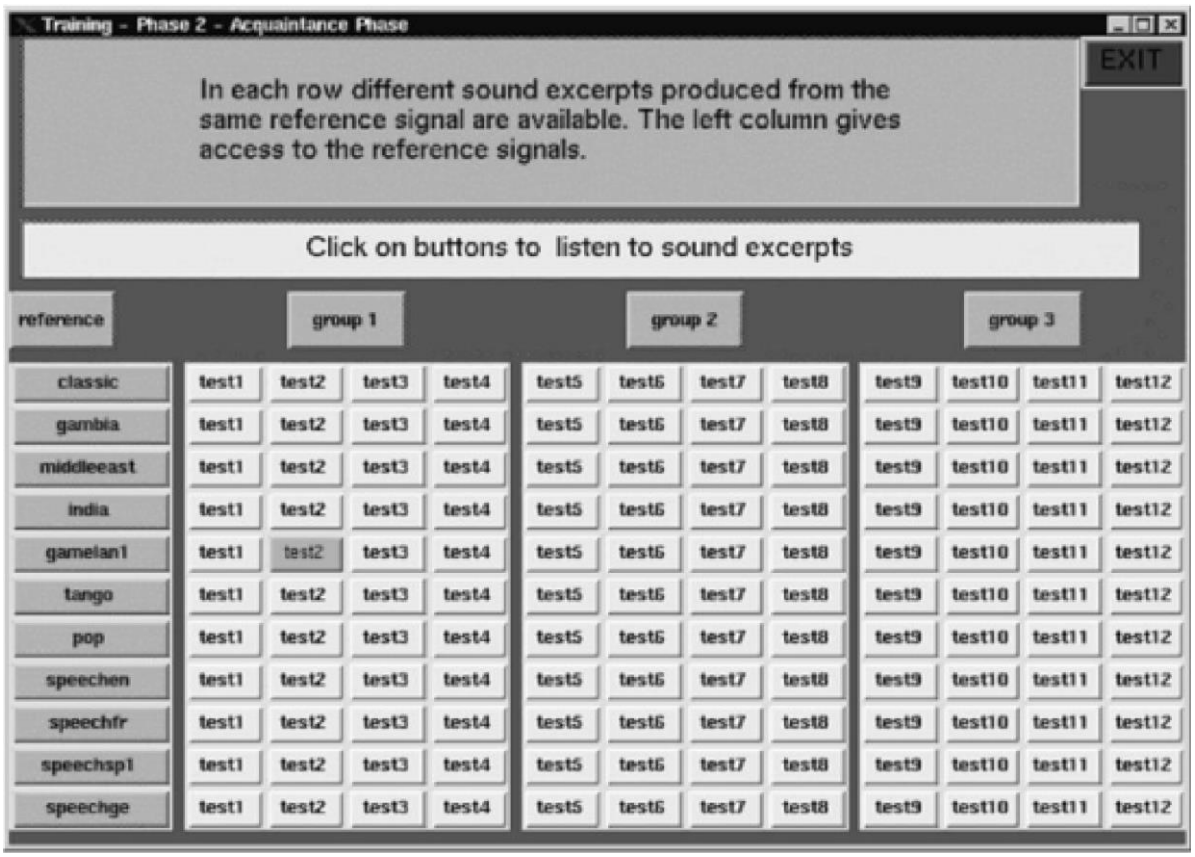
2 Этап слепой оценки

На этапе слепой оценки вам будет предложено выставить свои оценки по установленной шкале качества. Выставляемые вами оценки должны отражать ваши субъективные суждения об уровне качества каждого прослушанного звукового фрагмента. В ходе каждого испытания для оценки будет представлено 9 сигналов. Длительность каждого фрагмента – около 10 с. Вам следует прослушать эталонный сигнал, опорные сигналы и сигналы всех испытуемых систем, нажимая соответствующие кнопки. Прослушивать сигналы можно в любом порядке сколько угодно раз.

Выскажите свое мнение о качестве каждого сигнала с помощью сопоставленного сигналу бегунка. Окончательно определившись с оценками всех сигналов, нажмите кнопку "Зафиксировать оценки" внизу экрана.

РИСУНОК 3

Пример пользовательского интерфейса для стадии A этапа обучения



BS.1534-03

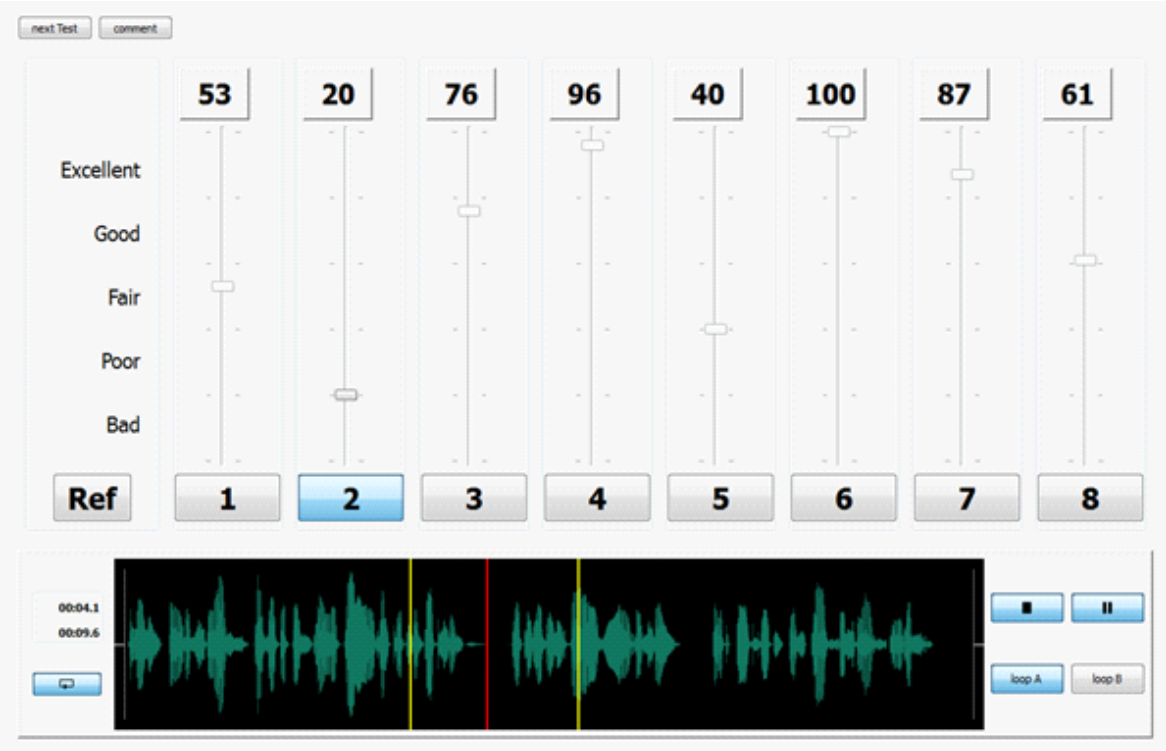
При выставлении оценок вы будете пользоваться шкалой качества, изображенной на рисунке 1.

Это непрерывная шкала, разбитая на категории от "отлично" до "очень плохо". Оценка 0 соответствует нижней границе категории "очень плохо", а оценка 100 – верхней границе категории "отлично".

При оценке звуковых фрагментов обратите внимание на то, что вы не обязаны давать оценку из категории "очень плохо" самому низкокачественному звуковому фрагменту в данном испытании. Вместе с тем один или несколько фрагментов должны обязательно получить оценку 100, поскольку один из оцениваемых звуковых фрагментов – это необработанный эталонный сигнал.

РИСУНОК 4

Пример пользовательского интерфейса для этапа слепой оценки



BS.1534-04

Прилагаемый документ 2
к Приложению 1
(информативное)

Руководящие указания по проектированию пользовательских интерфейсов

Приведенные ниже рекомендации адресованы тем, кто рассматривает возможность:

a) производства систем для выполнения субъективных испытаний по методу MUSHRA;

b) выполнения таких испытаний.

Эти рекомендации призваны повысить надежность результатов испытаний и облегчить анализ аномалий, выявленных при обработке оценок.

Пользовательский интерфейс должен быть спроектирован так, чтобы свести к минимуму вероятность выставления оценки, не соответствующей действительному намерению оценщика. Для этого следует позаботиться о том, чтобы из интерфейса оценщику было ясно, какую из обработанных версий тестового элемента он прослушивает в каждый момент времени. Здесь может помочь вдумчивый выбор цветов и уровней яркости экранных индикаторов (например, кнопок) для предупреждения возможных трудностей, связанных с нечувствительностью оценщика к некоторым цветам.

Следует также принять меры к тому, чтобы оценщик мог регулировать положение бегунка только для того элемента, который прослушивается в данный момент. По наблюдениям, некоторые оценщики последовательно прослушивают две обработанные версии звукового элемента с целью оценить первый, а не второй из них. В этих обстоятельствах есть вероятность ошибиться (особенно при большом количестве элементов управления на экране) и выставить оценку не тому сигналу, который имелся в виду. Чтобы по возможности снизить эту вероятность, рекомендуется в каждый момент времени оставлять активным только тот элемент управления, который связан с прослушиваемым в данный момент сигналом. Элементы управления, связанные с другими сигналами, следует на это время делать неработоспособными.

Прилагаемый документ 3
к Приложению 1
(нормативное)

Описание непараметрического статистического сравнения двух образцов с использованием методов перевыборки и численных методов Монте-Карло

Для определения значимости почти любого статистического результата можно использовать непараметрические критерии рандомизации в совокупности с общепринятыми методами перевыборки (например, бутстрапированием). Например, значимость наблюдаемой разницы в медианном отклике на два тестовых сигнала (с размерами выборок $N1$ и $N2$) можно определить следующим способом. Фиксируется фактическая разность между медианами выборок, которая обозначается как $Diff_{ACT\_1}$. Все данные из выборок собираются в единый файл или вектор. Выполняется процедура бутстрапирования, на каждой итерации которой производится перестановка совокупного множества со взятием выборок размерами $N1$ и $N2$ без замены. Далее фиксируется разность между медианами двух случайно взятых выборок, которая обозначается как $Diff_{EST\_1}$. Эта процедура затем повторяется 10 000 раз. Соответствующее значение p равно количеству раз, когда $Diff_{EST\_N}$ превысило $Diff_{ACT\_N}$, деленному на 10 000. Если общее количество раз, когда $Diff_{EST\_N}$ превысило $Diff_{ACT\_N}$, меньше 500 (500/10 000 = 0,05), разность между двумя средними может считаться значимой при уровне значимости 0,05, $p < 0{,}05$.

Прилагаемый документ 4
к Приложению 1
(информативное)

Руководящие указания по параметрическому статистическому анализу

1 Введение

Описание базовой процедуры параметрического статистического анализа результатов испытаний по методу MUSHRA приведено в разделе 9. Однако обобщенный метод анализа, такой как ANOVA, предпочтительнее множества попарных сравнений, особенно когда приходится сравнивать друг с другом множество различных условий. В настоящем приложении даются указания по выполнению такого анализа. Среди прочего описываются необходимые условия для анализа этого типа и предлагаются альтернативы для случаев, когда эти условия не выполняются.

В испытании по методу MUSHRA используется *план с повторяющимися измерениями* или *внутрисубъектный план* (великолепное введение в эти понятия можно найти в книге Максвелла и

Делейни, вышедшей в 2004 году), при которых два внутрисубъектных фактора (условие и аудиоматериал) полностью пересекаются, и каждому сочетанию слушателя, аудиоматериала и условия дается как минимум одна оценка. Возможны также случаи, когда одни и те же сочетания аудиоматериала и условия представляются вниманию двух или более групп оценщиков, например в разных лабораториях. В таких случаях присутствует дополнительный межсубъектный фактор "группа", который должен быть учтен при анализе.

Для обобщения результатов, полученных на относительно малой выборке из генеральной совокупности всех слушателей, необходимо обращение к статистике вывода. Например, если оценки, выставленные в ходе испытания с прослушиванием, указывают на разницу между воспринимаемым качеством звучания нового и традиционного кодера, важно ответить на вопрос, можно ли ожидать такую же разницу, если качество звучания будет оценивать совершенно другая группа слушателей. Что касается конкретного плана испытаний с прослушиванием по методу MUSHRA, имеет смысл дать ответы как минимум на три вопроса (или, на языке статистики, проверить три гипотезы), и описываемая здесь статистика вывода дает корректные ответы на эти вопросы. Наибольший интерес представляет вопрос о том, различалось ли воспринимаемое качество звучания испытуемых систем (например, эталонного сигнала и сигнала трех различных кодеров). Следующий вопрос таков: если бы в ходе испытания аудиосистемы оценивались на других тестовых материалах, зависели бы оценки качества звучания от аудиоматериала? Последний вопрос – различалось ли влияние аудиосистемы на воспринимаемое качество звука для разных тестовых материалов? Для надлежащего ответа на эти вопросы необходимо сначала проверить значимость основного эффекта условия (аудиосистемы), основного эффекта аудиоматериала и взаимодействия между условием и аудиоматериалом, выполнив дисперсионный анализ (по модели ANOVA). Взаимодействие присутствует, когда различия в воспринимаемом качестве звучания аудиосистем зависят от аудиоматериала. Обратите внимание, что ввиду потенциальных взаимодействий не рекомендуется агрегировать оценки каждой аудиосистемы на всем множестве аудиоматериалов, даже если эффекты аудиоматериала или взаимодействия не представляют особого интереса. После этого можно путем дополнительных сравнений проверить более конкретные гипотезы, касающиеся, например, воспринимаемой разницы между парой аудиосистем.

В случаях, когда стоит задача сравнить более двух условий эксперимента, например четыре различных кодера, основывать статистику вывода на множестве попарных сравнений некорректно. Например, если в испытании участвовали 5 аудиосистем ($K = 5$) (4 кодера и эталонный сигнал), это дает $\left(\frac{K}{2}\right) = K(K-1)/2 = 10$ пар условий. Если проверять различия отдельно в каждой из этих 10 пар по парному *t*-критерию при уровне значимости α, равном 0,05, это приведет к чрезмерному росту так называемой групповой вероятности ошибки первого рода. Для каждой отдельной проверки по *t*-критерию вероятность ошибочного отвержения нулевой гипотезы, заключающейся в отсутствии разницы в воспринимаемом качестве звучания двух кодеров, составляет α.

На множестве C таких проверок вероятность совершения хотя бы одной ошибки первого рода равна $1 - (1 - \alpha)^C$, что для $C = 10$, как в нашем примере, дает 0,40, что гораздо выше желаемого значения $\alpha = 0{,}05$. Снизить групповую вероятность ошибки можно путем соответствующих поправок на множественные проверки, например поправки Бонферрони или описываемой далее процедуры Хохберга (1988). Но даже и с учетом поправки попарные проверки по *t*-критерию будут маскировать представляющую интерес информацию – в частности потому, что такие множественные проверки на всех парах средних значений предполагают использование избыточной информации (каждое среднее участвует в нескольких проверках). Подход с попарными проверками характеризуется, как правило, меньшей мощностью (то есть меньшей чувствительностью к разнице между условиями), чем соответствующий обобщенный метод проверки, то есть дисперсионный анализ с повторяющимися измерениями (rmANOVA) в случае испытания по методу MUSHRA. Ниже дается пошаговое описание анализа данных испытания по методу MUSHRA, в котором отсутствуют факторы взаимодействия между участниками. Иными словами, предполагается, что в испытании участвовала только одна группа оценщиков и что все сочетания условий и аудиоматериала были представлены вниманию каждого оценщика хотя бы единожды. Обобщение на случай нескольких групп (например, когда испытание выполнялось в двух лабораториях) будет описано далее.

2 Проверка нормальности распределения

Благоразумным будет рассмотреть, как потенциальное отклонение распределения ответов от нормальности повлияет на корректность статистической проверки гипотез. В случае межсубъектного плана, когда вниманию каждого оценщика представляется только одно условие, проверки по методу ANOVA в рамках общей линейной модели удивительно робастны к ненормальности распределения ответов (например, [11], [13], [25], [35]).

В случае плана с повторяющимися измерениями, который используется в испытаниях по методу MUSHRA, прежде всего следует отметить альтернативный способ проверки нулевой гипотезы о том, что на генеральной совокупности воспринимаемое качество звучания идентично для всех условий. Это равносильно вычислению $K-1$ ортогональных контрастов, например путем определения разностных переменных для этих K условий с последующей проверкой гипотезы о том, что среднее значение этих разностных переменных по генеральной совокупности равняется 0. Например, в случае испытаний с эталонным сигналом и двумя кодерами можно для каждого участника определить две разностные переменные: D_1 – разность между оценками эталонного сигнала и кодера A, D_2 – разность между оценками кодера A и кодера B. Во всех видах анализа по методу ANOVA с повторяющимися измерениями предполагается, что эти разностные переменные имеют многомерное нормальное распределение. К сожалению, в отличие от межсубъектных планов здесь ненормальность может привести к излишне консервативным оценкам вероятности ошибок первого рода ([5], [22], [30], [39]). Это означает, что при заданном уровне значимости α (например, $\alpha = 0{,}05$) доля случаев, когда метод ANOVA дает значимое p-значение ($p < \alpha$) в условиях истинности нулевой гипотезы об идентичности средних для всех условий, будет меньше или больше номинального значения α. Опять-таки, в отличие от межсубъектных планов, эта проблема не решается одним только увеличением размера выборки [30]. Растет число свидетельств о том, что отклонения от симметрии имеют гораздо более серьезные последствия, чем отклонения от нормальности, характеризуемые куртозисом ([4], [18]). Степень отклонения от симметрии можно описать параметром *коэффициента асимметрии* распределения, который представляет собой центральный момент третьего порядка [8]. Для симметричного распределения, подобного нормальному, коэффициент асимметрии равен 0. *Куртозис* – центральный момент четвертого порядка, описывающий остроту пика и длину хвостов распределения (иллюстрации см. в [9]). Результаты ранее проведенных исследований по статистическому моделированию показывают, что для малых отклонений от симметрии методы rmANOVA сохраняют сдерживающий эффект в отношении вероятности ошибок первого рода. Но сегодняшнее состояние науки не позволяет сформулировать точные правила о приемлемой степени отклонения от нормальности. Поэтому важно проверить гипотезу о многомерном нормальном распределении и привести в протоколе эмпирические оценки коэффициента асимметрии и куртозиса.

Важно отметить, что общая линейная модель, лежащая в основе методов rmANOVA, не предполагает нормальности распределения исходных ответов (то есть оценок в испытании по методу MUSHRA). Вместо этого модель предполагает нормальное распределение *ошибок*. По этой причине проверка нормальности (вычисление коэффициента асимметрии и куртозиса) должна производиться для *остаточных ошибок* модели, а не для исходных данных. К счастью, в большинстве статистических программных пакетов предусмотрена возможность сохранять остаточные ошибки для каждого анализируемого условия эксперимента (в нашем случае – каждого сочетания аудиосистемы и аудиоматериала). Это дает один вектор остаточных ошибок для каждого условия эксперимента. В каждом векторе одно значение представляет одного оценщика.

Существует целый ряд критериев многомерного нормального распределения, например многомерный критерий Шапиро-Уилка, предложенный Ройстоном [34], критерии на основе многомерных оценок асимметрии и куртозиса [10], другие подходы [14]. Макросы для применения таких критериев имеются в SPSS (http://www.columbia.edu/~ld208/normtest.sps) и SAS (http://support.sas.com/kb/24/983.html), а также с большой вероятностью и в других пакетах. Одномерные оценки коэффициента асимметрии и куртозиса, которые могут быть отдельно рассчитаны для остаточных ошибок каждого сочетания аудиосистемы и аудиоматериала, предоставляются всеми основными статистическими программными пакетами. Макрос для SPSS авторства ДеКарло [9] (http://www.columbia.edu/~ld208/normtest.sps) обеспечивает также расчет многомерных коэффициента асимметрии и куртозиса [26]. Оценки одно- или многомерных коэффициента асимметрии и куртозиса, а также результат проверки гипотезы о многомерном нормальном распределении следует привести в протоколе.

Если подтверждение гипотезы о многомерном нормальном распределении не является значимым или ни один много- или одномерный критерий не показывает значимого отклонения коэффициента асимметрии или куртозиса от значений, ожидаемых для нормального распределения, то предположения метода rmANOVA выполняются.

Если же по какому-либо из критериев обнаруживается значимое отклонение от нормальности или если коэффициент асимметрии для любого условия эксперимента превышает 0,5 (в качестве предварительного практического ориентира), то встает вопрос о том, какие выводы следует из этого делать. Есть две общие проблемы, которые связаны с упоминавшимся выше отсутствием формальных правил касательно приемлемого отклонения от нормальности для методов rmANOVA. Во-первых, критерии многомерного нормального распределения довольно чувствительны и зачастую обнаруживают незначительные отклонения от нормальности. Вдобавок чувствительны они не только к асимметрии в распределении остаточных ошибок, но также и к куртозису и другим аспектам распределения, тогда как скорее всего только асимметрия приводит к необастным ошибкам первого рода в методах rmANOVA. Во-вторых, если оценивать многомерные коэффициент асимметрии и куртозис по данным [26], полученная информация не позволяет сделать вывод о возможности применения метода rmANOVA, что опять-таки обусловлено отсутствием правил касательно приемлемого отклонения от нормальности. Это подчеркивает необходимость указывать в протоколе коэффициент асимметрии и куртозис, а также результаты статистической проверки гипотез. Как только будут выяснены корректные правила определения приемлемого отклонения от нормальности, можно будет пересмотреть результаты проверки гипотез методом rmANOVA, имея более полную информацию. Если отклонение от нормальности представляется серьезным, на что указывает, в частности, оценка коэффициента асимметрии выше 1,0 [29], можно рассмотреть непараметрические альтернативы методу rmANOVA, например критерии с использованием методов перевыборки или критерий Фридмана. Однако пока еще неясно, в каких ситуациях методы перевыборки решают проблему ненормальности [38]. Критерий Фридмана не предполагает многомерного нормального распределения, зато предполагает равенство дисперсии для всех условий эксперимента [36], что часто не выполняется для экспериментальных данных. Сверх того критерий Фридмана является одномерным. Поэтому даже если предположение о равенстве дисперсий выполняется, критерий Фридмана можно использовать для выявления эффекта аудиосистемы, усредненного по всему множеству аудиоматериала, но нельзя – для анализа взаимодействия между аудиосистемой и аудиоматериалом.

3 Выбор подхода к анализу методом rmANOVA

Есть много подходов к статистической проверке эффектов внутри- и межсубъектных факторов в данных экспериментов с повторяющимися измерениями [21]. Поскольку сейчас мы рассматриваем случай эксперимента, в котором отсутствуют межсубъектные (группировочные) факторы, и предполагаем полный набор данных (то есть наличие оценок для каждого сочетания слушателя, аудиоматериала и условия), можно рекомендовать два подхода. Оба эти подхода предоставляют корректные критерии для проверки гипотез при многомерном нормальном распределении данных, но могут различаться по статистической мощности (то есть чувствительности к отклонению от нулевой гипотезы) в зависимости от других эффектов, в частности от размера выборки.

Это (a) *одномерный подход с поправкой Хюнха-Фельдта на степени свободы* и (b) *многомерный подход*. Подробные описания этих подходов можно найти в литературе [21], [28]. Оба варианта доступны в основных статистических программных пакетах (например, R, SAS, SPSS, Statistica).

Ввиду структуры данных, характерной для повторяющихся измерений, существует корреляция между оценками, полученными для разных сочетаний условия и аудиоматериала. Например, если слушатель выставляет необычно высокую оценку низкокачественному опорному сигналу, его/ее оценки кодеров будут, вероятно, также выше, чем у других оценщиков. Многомерный подход предполагает сферическую структуру дисперсии-ковариации данных, что эквивалентно равенству дисперсии описанных ниже разностных переменных [16], [33]. Но это предположение нарушается практически во всех эмпирических наборах данных [21]. Для решения этой проблемы при расчете *p*-значения для *F*-распределения к степеням свободы применяют поправочный коэффициент. При этом отклонение от сферичности оценивается по данным. Рекомендуется [17] использовать поправочный

коэффициент Хюнха-Фельдта, обозначаемый как $\tilde{\varepsilon}$, поскольку альтернативный поправочный коэффициент Гринхауза-Гейсера [12] имеет тенденцию давать консервативные оценки (например, [17], [30]). Когда данные имеют нормальное распределение, одномерный подход с поправкой Хюнха-Фельдта дает корректные вероятности ошибок первого рода даже на выборках чрезвычайно малого размера (N = 3). Поправочный коэффициент $\tilde{\varepsilon}$ и исправленные p-значения предоставляются всеми основными статистическими программными пакетами.

В *многомерном подходе* используется не идентичная, но эквивалентная формулировка нулевой гипотезы. Например, рассмотрим нулевую гипотезу о том, что на генеральной совокупности воспринимаемое качество звучания идентично для всех условий. Это равносильно вычислению $K - 1$ ортогональных контрастов, например путем определения разностных переменных для этих K условий с последующей проверкой гипотезы о том, что вектор μ средних значений по генеральной совокупности всех $K - 1$ контрастов равен нулевому вектору, $\mu = 0$. Например, в случае испытаний с эталонным сигналом и двумя кодерами можно для каждого участника определить две разностные переменные: D_1 – разность между оценками эталонного сигнала и кодера A, D_2 – разность между оценками кодера A и кодера B. Многомерный вариант метода rmANOVA основывается на разностных переменных и предусматривает многомерную проверку гипотезы $\mu = 0$. Этот подход не требует делать предположений о матрице дисперсии-ковариации. Для данных, имеющих многомерное нормальное распределение, этот критерий является точным, но требует привлекать как минимум столько оценщиков, сколько есть уровней фактора. Поэтому его нельзя использовать, например, если условий 9 (8 кодеров и эталонный сигнал), а оценщиков всего 8.

Относительная мощность двух подходов зависит, среди множества других факторов, от размера выборки и количества уровней внутрисубъектного фактора. Элджайна и Кесельман (1997 год) предлагают простое правило выбора: использовать одномерный подход с поправкой Хюнха-Фельдта, если $\tilde{\varepsilon} > 0{,}85$ и $N < K + 30$, где N – число оценщиков, а K – максимальное число уровней внутрисубъектных факторов. В остальных случаях следует использовать многомерный подход. Обратите внимание, что если эксперимент проводился в разных лабораториях, то N – общее число оценщиков, участвующих в исследовании (например, 10 оценщиков в лаборатории A и 10 оценщиков в лаборатории B дают N = 20).

4 Выполнение анализа по методу rmANOVA и необязательных апостериорных проверок

На этом шаге проверяются гипотезы об эффектах условия, аудиоматериала и их взаимодействий по обобщенным критериям в варианте rmANOVA. Для вычислений по методу rmANOVA в большинстве программных пакетов, таких как SAS, SPSS и Statistica, требуется, чтобы данные были представлены в форме "одна строка на одного оценщика". Соответственно таблица данных должна быть структурирована указанным образом, и оценки всех сочетаний условий и аудиоматериалов представляются в виде столбцов (переменных).

Двухфакторный анализ по методу rmANOVA предоставляет информацию о трех эффектах.

1) Основной эффект условия

В большинстве случаев эта проверка представляет наибольший интерес. Если анализ по методу ANOVA указывает на значимый эффект условия, то можно отвергнуть нулевую гипотезу о том, что на генеральной совокупности воспринимаемое качество звучания идентично для всех условий (эталонный сигнал и кодеры с 1 по k). Иными словами, эта проверка показывает, что на генеральной совокупности воспринимаемое качество звучания аудиосистем различается. В качестве характеристики величины эффекта нельзя использовать предложенный Коэном [6] показатель d или его аналоги, поскольку показатель d не определен для сравнения более чем двух средних значений. В контексте ANOVA обычно в протоколе характеризуют степень положительной связи. Эта характеристика показывает, какая доля дисперсии данных определяется рассматриваемым эффектом. Те же соображения лежат в основе коэффициента детерминации R^2. Большинство статистических программных пакетов позволяют вычислять частичный эмпирический коэффициент детерминации η^2, который равен отношению дисперсии, обусловленной рассматриваемым эффектом, к сумме этой дисперсии и дисперсии, обусловленной ошибками (остаточными). Обсуждение альтернативных показателей степени положительной связи можно найти в работе [Олейник и Элджайна] [31].

После того как проверкой установлена значимость основного эффекта, дальше нередко встает задача определить его происхождение. Выяснить его можно путем вычисления частных контрастов. Пусть, например, необходимо определить, есть ли разница в качестве звучания нового кодера и трех традиционных систем. Для ответа на этот вопрос сначала для каждого оценщика вычисляют усредненную по всему аудиоматериалу оценку трех традиционных кодеров. Таким образом, от каждого оценщика имеется a) одна оценка нового кодера и b) одна усредненная оценка трех прочих кодеров. Эти два значения затем сравниваются с применением парного *t*-критерия. Обратите внимание, что поскольку данные взяты из эксперимента с повторяющими измерениями, важно не использовать объединенную выборочную дисперсию [27]. Также следует иметь в виду, что этот контраст можно проверить как плановый вместо выполнения анализа по методу ANOVA. В общем случае рекомендуется пользоваться двусторонними критериями значимости. Но если, например, есть *априорная* гипотеза о том, что новый кодер должен получать более высокие оценки, чем традиционные кодеры, допустимо задавать одностороннюю область отвержения.

Другие частные контрасты можно вычислять, применяя те же рассуждения. Более общий рецепт проверки контрастов – вычислить линейную комбинацию оценок, полученных в разных условиях эксперимента, а затем с помощью парного *t*-критерия определить, значимо ли отличие этого контраста от нуля. Для каждого оценщика i значение контраста вычисляется по формуле

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

где Y_{ij} – оценка, выставленная оценщиком i для условия j (усредненная по всему аудиоматериалу), a – количество условий, учитываемых в этом контрасте, а c_j – коэффициенты. Если в приведенном выше примере новому кодеру соответствует $j = 1$, а трем остальным кодерам $j = 2…4$, то контраст с коэффициентами $c_1 = -1$ и $c_2 = c_3 = c_4 = 1/3$ позволит проверить гипотезу о том, что качество звучания нового кодера и трех остальных кодеров отличается.

При вычислении более одного апостериорного контраста возникают проблемы, связанные с множественными проверками, как обсуждалось выше. Для их решения рекомендуется применять предложенную Хохбергом [15] восходящую процедуру Бонферрони с последовательной проверкой гипотез. Эта процедура ограничивает групповую вероятность ошибок первого рода, отличаясь при этом большей мощностью от многих альтернативных процедур [20]. В процедуре Хохберга сначала вычисляют m представляющих интерес контрастов, которые затем упорядочивают по *p*-значению. Если после этого наибольшее *p*-значение меньше α, то все гипотезы отклоняют (то есть все контрасты значимы). Если нет, значит по *t*-критерию контраст с наибольшим *p*-значением оказался незначимым, и тогда переходят к сравнению следующего по величине *p*-значения с $\alpha/2$. Если *p*-значение меньше, то данный контраст и все контрасты с меньшими *p*-значениями являются значимыми. Если нет, значит и этот контраст незначим, и тогда переходят к сравнению следующего по величине *p*-значения с $\alpha/3$. Говоря более формальным языком, если p_i, где $i = m, m - 1, \ldots 1$, упорядоченные по убыванию *p*-значения, то для любого $i = m, m - 1, \ldots 1$ справедливо: если $p_i < \alpha/(m - i + 1)$, то все контрасты с $i' \leq i$ значимы.

Теоретически можно также проводить апостериорные попарные сравнения оценок для всех условий. В случае плана с повторяющимися измерениями для этого потребовалось бы проверить все пары условий по парному *t*-критерию. Но так поступать не рекомендуется. Рассмотрим эксперимент с 7 кодерами и одним опорным сигналом. Для этого набора из 8 условий понадобится $8 \cdot 7/2 = 28$ попарных сравнений, и из такого большого числа проверок нелегко будет извлечь осмысленную информацию. Если статистической проверке подлежат все внутрипарные различия, то очевидно, что ввиду большого числа проверок особенно важным будет применить поправку на множественные проверки по методу Хохберга [15]. Обратите внимание, что если есть свидетельства отклонения от нормальности разностных оценок, на которых основывается проверка по парному *t*-критерию, то существует альтернатива, не предполагающая нормальности, – знаковый критерий.

Следует отметить, что после установления значимости основного эффекта может случиться, что все апостериорные контрасты или внутрипарные различия окажутся незначимыми [28] из-за использования разной статистической информации в методе rmANOVA и в апостериорных проверках. Важно, что rmANOVA – более подходящий метод проверки. Поэтому эффект,

признанный значимым по методу ANOVA, остается таковым, даже если ни одна апостериорная проверка не показывает значимости. Если обобщенная проверка (ANOVA) показала значимость, но ни один апостериорный контраст не значим, то можно заключить, что воспринимаемое качество звучания аудиосистем различалось. Можно также сравнивать различия между аудиосистемами. Например, для пар аудиосистем с наибольшей разницей в оценках качества звучания высока вероятность, что эти внутрипарные различия окажутся значимыми при большем размере выборки. Необходимо однако заключить, что в настоящем исследовании ни одно из внутрипарных различий не было значимым.

Если проверка по методу rmANOVA *не обнаруживает* значимости основного эффекта условия, это свидетельствует о малости различий между испытуемыми системами. Однако ввиду конечного размера выборки нельзя сделать вывод, что на генеральной совокупности *отсутствуют* различия в воспринимаемом качестве звука между условиями [3]. Причиной может быть как отсутствие различий на генеральной совокупности, так и недостаточность величины эффекта для его обнаружения на выборке данного размера. Если проводился *априорный* анализ мощности, то есть размер выборки был взят достаточным для выявления эффекта заданной величины с заданной вероятностью, то можно заключить, что данные свидетельствуют против наличия эффекта указанной величины.

Этот вывод можно принять за определение прозрачности кодеров. Если же *априорный* анализ мощности не проводился, следует проявить осмотрительность в выводах о прозрачности кодеров по изложенным выше причинам. Обычное приблизительное апостериорное решение – сравнивать *p*-значение с [0,2], а не 0,05. Если проверка по-прежнему не показывает значимости, это служит несколько более сильным указанием на отсутствие различий в воспринимаемом качестве звучания аудиосистем.

2) Основной эффект аудиоматериала

Действуя в том же порядке и руководствуясь теми же соображениями, можно проверить значимость основного эффекта аудиоматериала, что даст информацию о систематических отклонениях оценок, обусловленных тестовым материалом. В большинстве испытаний по методу MUSHRA этот эффект не должен представлять большого интереса, поскольку он не связан с различиями между аудиосистемами.

3) Взаимодействие условия и аудиоматериала

Если анализ по методу rmANOVA обнаруживает значимое взаимодействие между условием и аудиоматериалом, то аудиосистема по-разному влияет на воспринимаемое качество звучания в зависимости от тестового материала. Например, в случае эстрадной песни с высоким уровнем компрессии, где артефакты кодирования маскируются изначально присутствующими в материале искажениями, эталонный сигнал и кодер могут получить одинаковые оценки. С другой стороны, на записи концертного рояля с высоким динамическим диапазоном кодер может получить более низкую оценку качества звучания по сравнению с эталонным сигналом. Это взаимодействие, как правило, представляет интерес в испытании по методу MUSHRA, поскольку оно указывает на то, что воспринимаемая разница между аудиосистемами зависит от тестового материала.

Если по итогам обобщенной проверки эффект взаимодействия признан значимым, можно далее исследовать характер этого взаимодействия путем апостериорных проверок. Распространенный подход – проверка значимости *простых основных эффектов*. Ее можно провести, например, выполнив несколько отдельных однофакторных проверок по методу rmANOVA с внутрисубъектным фактором, по одной на каждый аудиоматериал. Эти проверки покажут, для каких аудиоматериалов наблюдался значимый эффект условия. Здесь также следует внести поправку на множественные проверки, используя процедуру Хохберга.

Как и прежде все внутрипарные различия между сочетаниями условия и аудиоматериала теоретически можно проверить отдельно по парному *t*-критерию с процедурой Хохберга. Однако число попарных сравнений в этом случае окажется еще большим, чем для основных эффектов. Например, 8 аудиосистем и 4 тестовых материала дают 24 сочетания, что соответствует $24 \cdot 23/2 = 276$ попарным сравнениям. Очевидно, что рекомендовать такой подход нельзя.

5 Обобщение на планы экспериментов с межсубъектной (группировочной) переменной

До сих пор мы рассматривали план эксперимента, в котором отсутствовали межсубъектные факторы. Какого рода анализ следует проводить, если в испытании участвовали разные группы оценщиков (например, из двух лабораторий или музыканты и немузыканты)?

При наличии межсубъектных факторов чрезвычайно важно, было ли число оценщиков одинаковым во всех группах (сбалансированный план) или разным (несбалансированный план).

Сбалансированный план. Если число оценщиков было одинаковым для всех уровней межсубъектного фактора или если размеры групп отличались не более чем на 10%, то при выполнении анализа по методу rmANOVA можно вновь применить одномерный подход с поправкой Хюнха-Фельдта на степени свободы или многомерный подход [21]. В дополнение к внутрисубъектным факторам условия и аудиоматериала план эксперимента будет теперь содержать как минимум один межсубъектный фактор (например, лабораторию). Поэтому анализ по методу rmANOVA будет предусматривать дополнительную проверку межсубъектных эффектов, а также взаимодействий между всеми внутри- и межсубъектными эффектами.

Например, может обнаружиться значимое взаимодействие между условием и лабораторией, то есть различия в воспринимаемом качестве аудиосистем между лабораториями A и B. Обратите внимание: здесь предполагается, что вниманию всех групп были представлены идентичные сочетания условия и аудиоматериала. Если, например, в двух лабораториях оценивались разные аудиоматериалы, то предлагаемые здесь методы использовать нельзя. Вместо этого потребуются так называемые модели случайных эффектов [28], рассмотрение которых выходит за рамки настоящего Приложения.

Несбалансированный план. Если размеры групп отличались более чем на 10%, то, к сожалению, ни одномерный, ни многомерный подходы уже не дадут корректных результатов [21]. Поэтому настоятельно рекомендуется предусмотреть в плане одинаковые размеры групп во избежание этой проблемы. Для групп существенно разного размера можно рекомендовать две процедуры анализа – усовершенствованную общую проверку с аппроксимацией (IGA) [1] и частную разновидность анализа по смешанной модели на основе метода максимального правдоподобия [23]. Для проверки IGA существует специальный макрос в программном пакете SAS. Анализ по смешанной модели можно выполнять, например, в программном обеспечении SAS PROC MIXED. Для последнего анализа важно установить два параметра. Во-первых, необходимо рассчитать степени свободы по методу, описанному в [19]. Для осуществления этого с использованием пакета SAS необходимо в операторе `model` задать параметр `ddfm=KR`. Во-вторых, необходимо аппроксимировать гетерогенную межсубъектную неструктурированную структуру ковариации [проверить] (UN-H) [23], задав в операторе repeated параметры type=UN group=groupingvar, где groupingvar – имя переменной, содержащей классификацию группы.

Справочные документы

[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 and 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.

Прилагаемый документ 5
к Приложению 1
(информативное)

Требования к оптимальному поведению опорных сигналов

Ниже приведены важнейшие атрибуты, к оптимальному воплощению которых необходимо стремиться при проектировании любого результативного опорного сигнала.

Оптимальное поведение опорного сигнала должно:

1) продуцировать данные, которые существенно не меняют относительного порядка испытуемых систем в сравнении с данными, собранными с использованием опорных сигналов, определенных в Рекомендации МСЭ-R BS.1534;

2) способствовать выставлению субъективных оценок в более широком диапазоне шкалы качества испытуемых систем по сравнению с данными, полученными по испытуемым системам с использованием опорных сигналов, определенных в Рекомендации МСЭ-R BS.1534;

3) восприниматься слушателями как более похожее на сигналы испытуемых систем, чем опорные сигналы, определенные в Рекомендации МСЭ-R BS.1534. Это может в свою очередь приводить к увеличению времени оценки опорного сигнала;

4) обеспечивать достаточную чувствительность при сравнении испытуемых систем из среднего диапазона качества;

5) приводить к разнице приблизительно в 20–30 баллов между оценками опорных сигналов нижнего и среднего диапазона;

6) приводить к ухудшению качества опорных сигналов, лишь ограниченно зависящему от контента.

_____