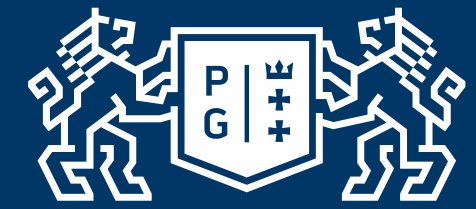


POLITECHNIKA
GDAŃSKA

Pomiary w Technice Studyjnej:

przeprowadzanie i analiza wyników testów subiektywnych

mgr inż. Adam Kurowski

Przeprowadzanie i analiza wyników testów subiektywnych

Wprowadzenie

Testy subiektywne to szczególny rodzaj badania, w którym **stwierdza się jak obiektywnie mierzalny czynnik wpływa na odczucia osób biorących udział w badaniu** (tzw. ekspertów).

Jest to metoda pozwalająca na **zobiektywizowanie badań takich właściwości jak jakość dźwięku lub obrazu** (np. dostarczana przez nowy kodek multimedialny).

Testy te **mają zazwyczaj formę swego rodzaju ankiety**, w której uczestnicy dokonują wyboru pomiędzy różnymi wersjami badanego materiału. W **zależności od badanego parametru zmianom mogą podlegać np. sposób prezentacji materiału, skala w jakiej udzielana jest odpowiedź, czy też technika zebranych w ten sposób danych**.

Dane zebrane w trakcie testów subiektywnych **muszą być poddane obróbce za pomocą metod wnioskowania statystycznego, aby otrzymane wyniki były wiarygodne**.

Standardy przeprowadzania testów subiektywnych

W zależności od potrzeb, **konstrukcja testu subiektywnego może być modyfikowana**.

Przykładami specyficznych potrzeb jest na przykład konieczność **porównania jakości sygnałów fonicznych** w przypadku, **gdy towarzyszy im jednocześnie obraz**.

Ocena jakości sygnałów fonicznych jest często wymagana na przykład przy **opracowywaniu nowych kodeków audio**. Może być też wykonywana w innych sytuacjach na przykład, aby **ocenić czy dany styl miksowania muzyki jest bardziej lub mniej preferowany przez daną grupę słuchaczy**.

Ze względu na częstą konieczność oceny jakości rozwiązań technicznych, **sposoby przeprowadzania testów subiektywnych są standaryzowane przez Międzynarodową Unię Telekonunikacyjną (ITU).** Zalecenia zawarte są w publikowanych przez ITU rekomendacjach.

Dobór metody testowania według ITU

W zaleceniach sformułowanych przez ITU można znaleźć szereg **wskazówek i wytycznych** dotyczących nie tylko **sposobu przeprowadzania** poszczególnych rodzajów testów subiektywnych, ale też **i ich doboru**.

Rodzaj testu subiektywnego zależy od kontekstu, tj. od tego **jaki sygnał** jest wykorzystany w teście (np. mowa, muzyka, obraz ruchomy, itp.) i **własności** którą mierzymy (np. lekka lub silna degradacja sygnału).

W zależności od tego, **jakie sygnały oceniamy zmieniać mogą się wymagania odnośnie takich rzeczy jak typ słuchaczy** oceniających sygnały, czy **sposób zadawania pytań** w samym teście.

Wytyczne dotyczące wyboru metody zawarte są w rekomendacji **ITU-R BS.1283-2**.

Dobór i przygotowanie słuchaczy według rekomendacji ITU

Pierwszym **kryterium** doboru słuchaczy jest ich **doświadczenie** w analitycznym słuchaniu nagrań i **umiejętność spostrzegania zniekształceń**. Osoby takie określane są mianem **ekspertów**

W zależności od celu testu **możliwe jest także wykorzystanie odpowiedzi od osób niemających takiego doświadczenia**. **Zależy to jednak od celu** przeprowadzania testu odsłuchowego.

Osoby **niebędące ekspertami** mogą brać udział **w badaniach mających za zadanie określić, jak dane rozwiązanie będzie odbierać „przeciętny konsument”**, który nie jest wyszkolony w umiejętności spostrzegania badanych cech.

W eksperymentach w których oceniane są niewielkie różnice sygnałów zazwyczaj konieczne jest zebranie grona ekspertów, którzy mają duże doświadczenie w danej dziedzinie.

Dobór i przygotowanie słuchaczy według rekomendacji ITU

Nie każda rekomendacja określa minimalną liczbę próby dla eksperymentu. Często zalecane jest wykonanie odrębnej analizy statystycznej w celu określenia jej dokładnej wartości.

Część zaleceń podaje bardziej konkretne wartości, przykładowo rekomendacja ITU-R BS.1284 zawierająca ogólne wytyczne przeprowadzania testów subiektywnych zaleca:

- próbę zebraną z grupy o liczebności **minimum 10 osób** w przypadku, gdy **składa się tylko z ekspertów**,
- próbę zebraną z grupy o liczebności **minimum 20 osób** w przypadku, gdy składa się **ona tylko z osób niebędących ekspertami**.

Niezależnie od doświadczenia każda osoba biorąca udział w teście **przed przeprowadzeniem właściwej procedury eksperymentalnej powinna być przeszkolona.**

Dobór i przygotowanie słuchaczy według rekomendacji ITU

Przeszkolenie powinno obejmować:

- zaznajomienie **z zasadami prezentacji** próbek materiału
- zaznajomienie z **zasadami udzielania odpowiedzi** (skala, interpretacja skali, sposób odpowiadania),

Możliwa jest też prezentacja przykładowych nagrań. Mogą one ilustrować na przykład kolejne stopnie degradacji sygnału fonicznego, której wpływ na jakość nagrania będzie w danym teście badana.

Dodatkowo **mogą być stosowane statystyczne techniki pozwalające na weryfikację wiarygodności odpowiedzi** udzielanych przez osoby biorące udział w teście.

Możliwe jest dokonywanie takich sprawdzeń przed samym testem lub na podstawie analizy wyników już po jego przeprowadzeniu.

Kryteria doboru rozmiaru próby

Zalecenia ITU dostarczają zgrubnej informacji o tym, jaką próbę należy dobrać. **Czasami jednak chcielibyśmy mieć większa pewność na przykład co jakości oszacowania średniej oceny danego sygnału fonicznego.**

W takim przypadku można posłużyć się **Ogólne kryteria doboru rozmiaru próby dla wartości średniej** pod warunkiem, że wiemy jaką **wariancję (lub odchylenie standardowe)** mają oceny, lub pod warunkiem, że znamy tej **estymatę wariancji (odchylenia standardowego).**

W każdym przypadku zakładamy, że **nie chcemy popełnić błędu większego niż d**, wartość ta **stanowi połowę długości pożądanego przez nas przedziału ufności dla mierzonego parametru.**

Jako, że oszacowanie **bazuje na przedziałach ufności, konieczne jest określenie pożądanego poziomu ufności α.**

Ogólne kryteria doboru rozmiaru próby dla wartości średniej

Przy założeniu, że **znamy teoretyczne odchylenie standardowe** σ rozkładu wartości badanego parametru, rozmiar próby dany jest wzorem:

$$N = \frac{u_{1-\alpha/2}^2 \cdot \sigma^2}{d^2},$$

gdzie $u_{1-\alpha/2}^2$ jest odczytaną z tabeli wartością krytyczną dla testu dwustronnego bazującego na **rozkładzie Gaussa**, która odpowiada wybranemu **poziomowi istotności** α.

źródło: Górecki T., Podstawy statystyki z przykładami w R, BTC, 2011.

Ogólne kryteria doboru rozmiaru próby dla wartości średniej

Jeżeli **nie znamy odchylenia standardowego**, tylko **szacujemy je bezpośrednio z danych, które posiadamy**, to jesteśmy zmuszeni posługiwać się **rozkładem zmiennej t**, znaną przez nas **estymatę odchylenia standardowego oznaczamy przez s**:

$$N = \frac{t^2_{1-\frac{\alpha}{2},n_0-1} \cdot s^2}{d^2},$$

gdzie $t^2_{1-\frac{\alpha}{2},n_0-1}$ jest odczytaną z tabeli wartością krytyczną dla testu dwustronnego bazującego na rozkładzie **zmiennej t**, która odpowiada wybranemu **poziomowi istotności α**.

Wartość $n_0 - 1$ nazywamy **liczbą stopni swobody**, a n_0 to **liczba obserwacji na podstawie której oszacowaliśmy s**.

źródło: Górecki T., Podstawy statystyki z przykładami w R, BTC, 2011.

Czynności przygotowawcze przed wykonaniem testu

Pierwszą i zwykle najbardziej kosztowną procedurą jest zebranie grupy ekspertów i **przygotowanie miejsca**, w którym test ma być przeprowadzony.

Drugim ważnym etapem **przygotowawczym jest obróbka i przygotowanie samego materiału** będącego przedmiotem badania.

Konieczne jest **sformułowanie atrybutów**, **które będą badane**. Na tej podstawie **przygotowywane jest pytanie**, z którego **jasno powinno wynikać, co dana osoba ma oceniać i jak ta ocena przekłada się na odpowiedź (np. w postaci podania numeru w skali od 1-5).**

Stąd też **nie powinno się na raz przeprowadzać oceny zbyt wielu atrybutów** (najlepiej oceniać tylko jeden).

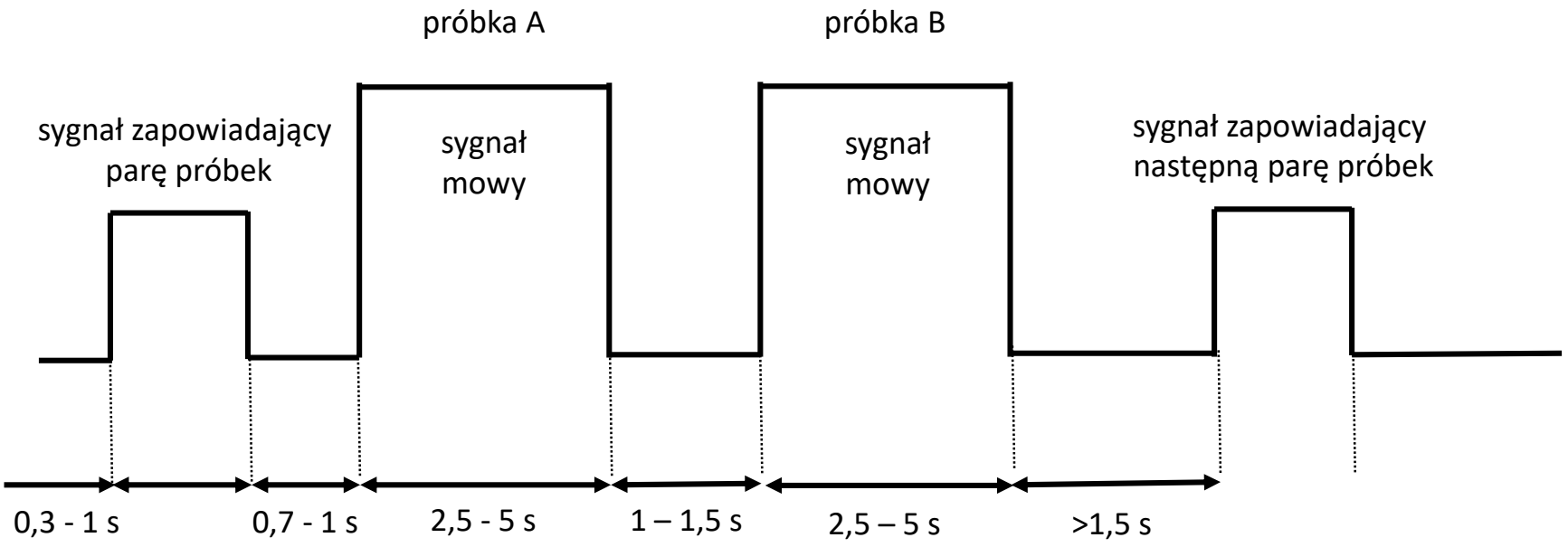
Czynności przygotowawcze przed wykonaniem testu

- Zaleca się by próbki prezentowanego materiału miały **długość od 2,5 do 5 sekund w przypadku sygnału mowy**.
- Zaleca się by próbki prezentowanego materiału miały **długość od 10 do 15 sekund w przypadku muzyki**.
- **Całość sesji** w trakcie której przeprowadzany jest **pojedynczy test** powinna trwać **od 20 minut do maksymalnie 45 minut**.
- Jeżeli **eksperyment wymaga bardzo długotrwałych pomiarów**, to można go **rozbić na kilka sesji, które powinny być rozdzielane przerwami.**

źródło: Rekomendacja ITU-T P.800.

Przeprowadzanie i analiza wyników testów subiektywnych

Czynności przygotowawcze przed wykonaniem testu



Przykładowa struktura prezentacji sygnału mowy proponowana w rekomendacji ITU-T P.800.

Zewnętrzny czynniki wpływające na wyniki testu

Zwłaszcza w przypadku testów odsłuchowych sygnałów o **niewielkich różnicach** w jakości konieczne jest zapewnienie **wysokiej klasy sprzętu**, na których odtwarzane są prezentowane próbki.

W przypadku stosowania **słuchawek** często preferowany jest **odsłuch binauralny**.

W przypadku **odsłuchu na monitorach referencyjnych** konieczne jest **upewnienie się, że na wyniki ocen słuchaczy nie mają wpływu takie czynniki jak ich umiejscowienie na sali**.

Wpływ akustyki pomieszczenia także powinien być znany. Przykładowo wg. Normy ITU-T P.800 czas pogłosu w pomieszczeniu w którym odbywa się test nie powinien przekraczać 300 ms w paśmie 125Hz – 8 kHz.

Dokumentacja prowadzenia i wyników testu subiektywnego

W **raporcie dokumentującym test** odsłuchowy zgodny z rekomendacjami ITU powinny być zawarte takie informacje jak:

- dane **o grupie osób** biorących udział w teście i **o wybranych fragmentach materiału** prezentowanego słuchaczom
- informacje **o miejscu** w którym przeprowadzany był test odsłuchowy takie jak **wymiary** pomieszczenia, jego **własności akustyczne** (np. czas pogłosu), **specyfikacje elektroakustycznych urządzeń** odtwarzających sygnały itp.,
- **opis procedury eksperymentalnej**, tj. procesu przygotowania i treningu słuchaczy, zasady prezentacji próbek, procedur testujących wiarygodność odpowiedzi itp..
- **wykorzystane metody analizy** danych pozyskanych z eksperymentu,
- **spis wniosków** wyciągniętych na podstawie wyników analiz z punktu poprzedniego.

źródło: Rekomendacja ITU-R BS.1284.

Przeprowadzanie i analiza wyników testów subiektywnych

Rekomendacja ITU-T P.800 – testowanie jakości mowy w telefonii

Oryginalny tytuł rekomendacji: SERIES P: TELEPHONE TRANSMISSION QUALITY - Methods for objective and subjective assessment of quality

Norma opisuje metodykę pomiaru jakości sygnału w sytuacji transmisji mowy lub muzyki przez łącze telefoniczne.

Wprowadza ona między innymi powszechnie używane **skale ocen bezwzględnych** bazujące na uśrednionej ocenie od 1 do 5 –MOS, $\mathrm{MOS}_{\mathrm{LE}}$, $\mathrm{MOS}_{\mathrm{LP}}$ (ang. mean opinion score).

Są to techniki **oceny jakości subiektywnej, w których nie jest podany punkt odniesienia**, konieczne jest doświadczenie eksperckie, aby wyniki były wiarygodne.

Dodatkowo definiuje możliwość oceny we **względnej, dwukierunkowej skali**, której wynikiem jest ocena CMOS.

Przeprowadzanie i analiza wyników testów subiektywnych

Rekomendacja ITU-T P.800 – testowanie jakości mowy w telefonii

Score	Listening-quality scale (MOS)	Listening-effort scale (MOS_{LE})	Loudness-preference scale (MOS_{LP})
5	Excellent	Complete relaxation possible; no effort required	Much louder than preferred
4	Good	Attention necessary; no appreciable effort required	Louder than preferred
3	Fair	Moderate effort required	Preferred
2	Poor	Considerable effort required	Quieter than preferred
1	Bad	No meaning understood with any feasible effort	Much quieter than preferred

Opis znaczenia ocen bezwzględnych o wartościach od 1-5 wg. Normy P.800.

Rekomendacja ITU-T P.800 – testowanie jakości mowy w telefonii

Dodatkowo rekomendacja definiuje procedurę pomiarową, w której porównywane są względem siebie dwie próbki (A i B), w takim przypadku wykorzystywana jest skala dwukierunkowa.

Uśredniona ocena uzyskana w wyniku oceny według takiej skali w normie ITU-T P.800 określana jest jako CMOS (ang. comparison mean opinion score).

Ze względu na istnienie sygnału odniesienia w tego typu teście możliwe jest uczestnictwo osób niebędących ekspertami.

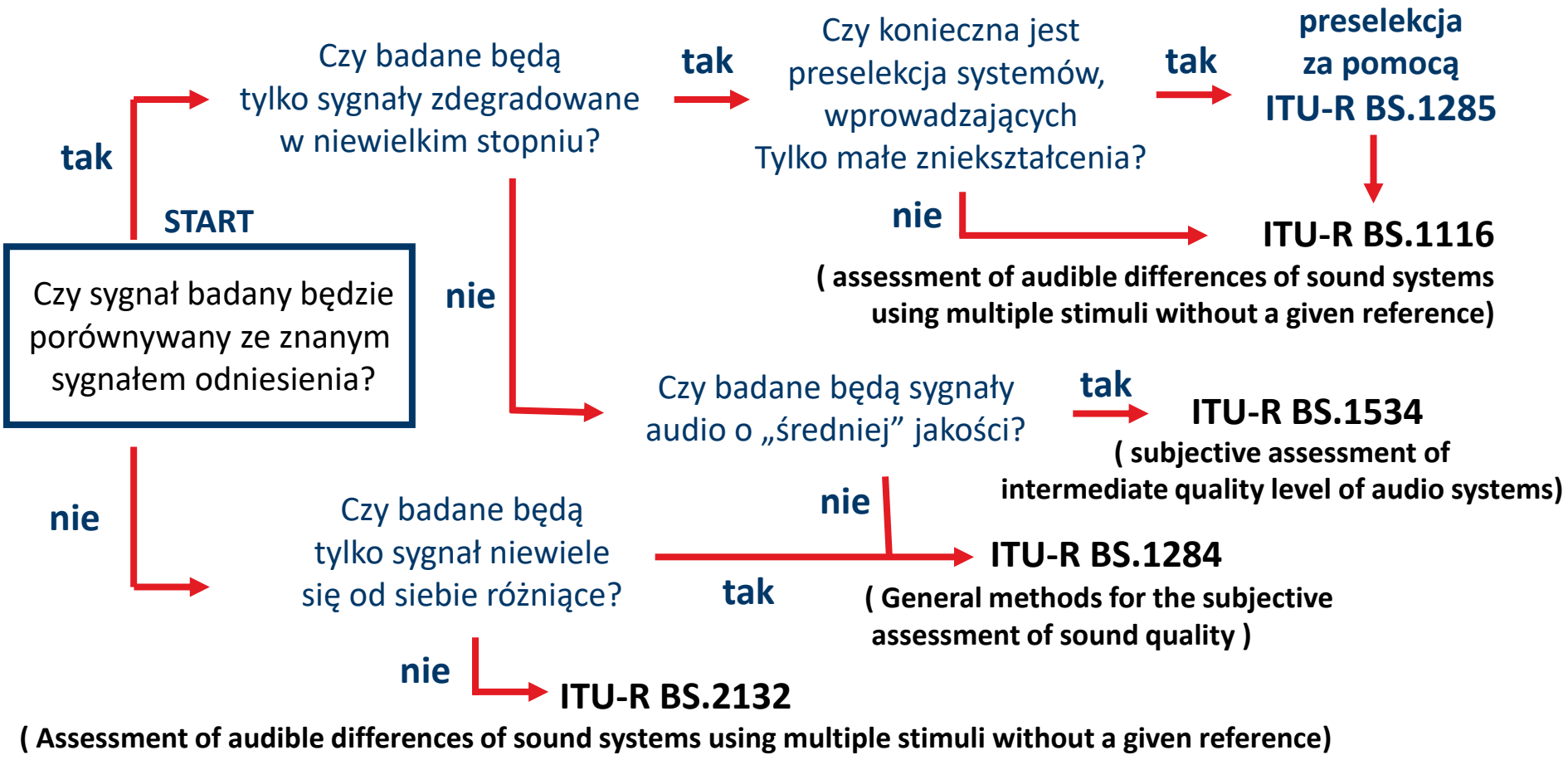
The Quality of the Second Compared to the Quality of the First is:

Score	Meaning
3	Much Better
2	Better
1	Slightly Better
0	About the Same
-1	Slightly Worse
-2	Worse
-3	Much Worse

POLITECHNIKA GDAŃSKA

Przeprowadzanie i analiza wyników testów subiektywnych

Dobór metody testowania według ITU-R (bez obrazu towarzyszącego)



Dobór metody testowania według ITU-R (z obrazem towarzyszącym)

Według rekomendacji **ITU-R BS.1283-2**, gdy badanie dotyczy oceny sygnałów akustycznych wraz z towarzyszącym obrazem nadal obowiązuje **diagram doboru jak dla braku obrazu towarzyszącego**.

Dodatkowo oprócz dobranej na tej zasadzie rekomendacji dodatkowo **obowiązują** zasady zawarte w rekomendacji **ITU-R BS.2126 (Methods for the subjective assessment of sound systems with accompanying picture)**.

Dobór metody testowania według ITU-R

W swoich rekomendacjach ITU rozróżnia **jakość sygnału (quality)** i tak zwaną **degradację sygnału (impairment)**.

Jakość jest subiektywnym odczuciem, **często mierzonym w skali bezwzględnej**.

Degradację zazwyczaj mierzy się **względem nagrania wzorcowego**, który nie jest zdegradowany. Zakłada się że jest to nagranie o najwyższej jakości w całym zbiorze prezentowanego materiału.

Pozostały materiał posiada mniejszą jakość w wyniku np. przetworzenia kodekiem stratnym.

Możliwe jest także **definiowanie własnych atrybutów**.

Przeprowadzanie i analiza wyników testów subiektywnych

Rekomendacja ITU-R BS.1284

Podtytuł rekomendacji: General methods for the subjective assessment of sound quality

Opisuje ogólne techniki przeprowadzania testów odsłuchowych z lub bez sygnału odniesienia.

Dodatkowo wprowadza skale o mniejszej gradacji ocen (od 0 do 100, lub od -60 do 60).

Wprowadza także interpretację skali 5-stopniowej pod względem oceny jakości (ang. **quality**) lub zniekształcenia (ang. **impairment**).

Wprowadzane są także zasady dokonywania porównań pomiędzy wieloma nagraniami.

Przeprowadzanie i analiza wyników testów subiektywnych

Rekomendacja ITU-R BS.1284

Score	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible, but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Opis znaczenia ocen bezwzględnych o wartościach od 1-5 wg. normy ITU-R BS.128.

Rekomendacja ITU-R BS.1116

Podtytuł rekomendacji: Methods for the subjective assessment of small impairments in audio systems

Jest to rekomendacja opisująca sposób przeprowadzania testów odsłuchowych w przypadku, gdy zniekształcenia są tak małe, że byłyby niewykrywalne bez rygorystycznego podejścia do prezentacji próbek ekspertom i odpowiedniej analizy statystycznej wyników tych testów.

Rekomendacja wprowadza między innymi test ABC, który określony jest jako test „podwójnie ślepy”, oparty na trójkach sygnałów, wśród których ukryty jest sygnał referencyjny

Przeprowadzanie i analiza wyników testów subiektywnych

Rekomendacja ITU-R BS.1116

W formule zawartej w rekomendacji osobom biorącym udział w teście prezentowane są trzy sygnały.

- **Sygnał A** – który **zawsze** jest sygnałem **referencyjnym** i jest to fakt znany osobie biorącej udział w teście odsłuchowym,
- **Sygnał B** – może być to albo powtórzony sygnał referencyjny, albo sygnał badany (zniekształcony),
- **Sygnał C** – jeśli sygnał B był sygnałem referencyjnym, to sygnał C zawiera sygnał badany, w odwrotnym przypadku – powtórzenie sygnału referencyjnego.

Słuchacz udziela **odpowiedzi w postaci swojej oceny zdegradowania sygnałów B i C względem sygnału A.**

Odpowiedzi są **wiarygodne, jeżeli słuchacz nie słyszy zdegradowania w parach referencja-referencja** (spodziewamy się opinii o braku różnic między sygnałami), a **jedynie istotnie statystycznie wskazuje na zdegradowanie sygnału niebędącego referencją**.

Rekomendacja ITU-R BS.1534

Podtytuł rekomendacji: Method for the subjective assessment of intermediate quality level of audio systems

Jest to rekomendacja wprowadzająca często stosowaną stosowaną formułę testu odsłuchowego, tzw. MUSHRA (ang. Multi Stimulus test with Hidden Reference and Anchor).

Jest to metoda oceny sygnałów średnio zniekształconych. Takich, które byłyby uznane za zbyt silnie zdegradowane w przypadku zastosowania technik z innych zaleceń (np. ITU-R BS.1116). Dla tego typu sygnałów w metodzie ABC wszystkie sygnały skumulowane są w dolnej części skali, konieczna jest zmiana metodyki.

Przeprowadzanie i analiza wyników testów subiektywnych

Rekomendacja ITU-R BS.1534

Sygnały w teście **MUSHRA oceniane są w skali od 0 do 100** (tzw. CQS – continuous quality scale).

Oprócz sygnałów podawanych ocenie dodatkowo zamieszczone są dwa sygnały specjalne:

- **sygnał referencyjny**, stanowiący sygnał odniesienia, niezawierający żadnych zniekształceń powodujących degradację jakości,
- tzw. **kotwica (ang. anchor), czyli sygnał o bardzo silnie zdegradowanej jakości**, jest to **degradacja wprowadzona celowo**, aby stanowił on **punkt odniesienia jako sygnał o najgorszej jakości** z całego zbioru materiału prezentowanego słuchaczom.

Każdy **sygnał może być odsłuchiwany wielokrotnie** przez osobę biorącą udział w teście.

Nie ma wymuszonej kolejności prezentacji nagrań.

Ze względu na dowolną możliwość odsłuchiwania przykładów bardzo **często test ten jest implementowany w postaci programu komputerowego**.

Przeprowadzanie i analiza wyników testów subiektywnych

Rekomendacja ITU-R BS.1534

Sygnałów **kotwicy może być kilka**, **najczęściej** występują **dwa** takie sygnały (tzw. **low anchor** i **medium anchor**).

Sygnnały w interfejsie osoby oceniającej ułożone są w losowej kolejności, słuchacz nie wie, które oceniane przez niego sygnały stanowią odniesienie, a które są faktycznie oceniane w teście.

Zaleca się, żeby test nie posiadał więcej niż 12 ocenianych fragmentów materiału:

- 9 sygnałów ocenianych,
- 1 sygnał kotwicy zniekształcony w silnym stopniu (ang. low anchor),
- 1 sygnał kotwicy zniekształcony w średnim stopniu (ang. mid anchor),
- 1 ukryty, niekształcony sygnał referencyjny.

Interfejs graficzny oprogramowania do przeprowadzania testu MUSHRA.

źródło: Rekomendacja ITU-R BS.1534.

Rekomendacja ITU-R BS.2132

Podtytuł rekomendacji: Method for the subjective quality assessment of audible differences of sound systems using multiple stimuli without a given reference

Jest to rekomendacja opisująca sposób przeprowadzania testów odsłuchowych **dla sygnałów różniących się w sposób znaczny w sposób** podobny do testu MUSHRA.

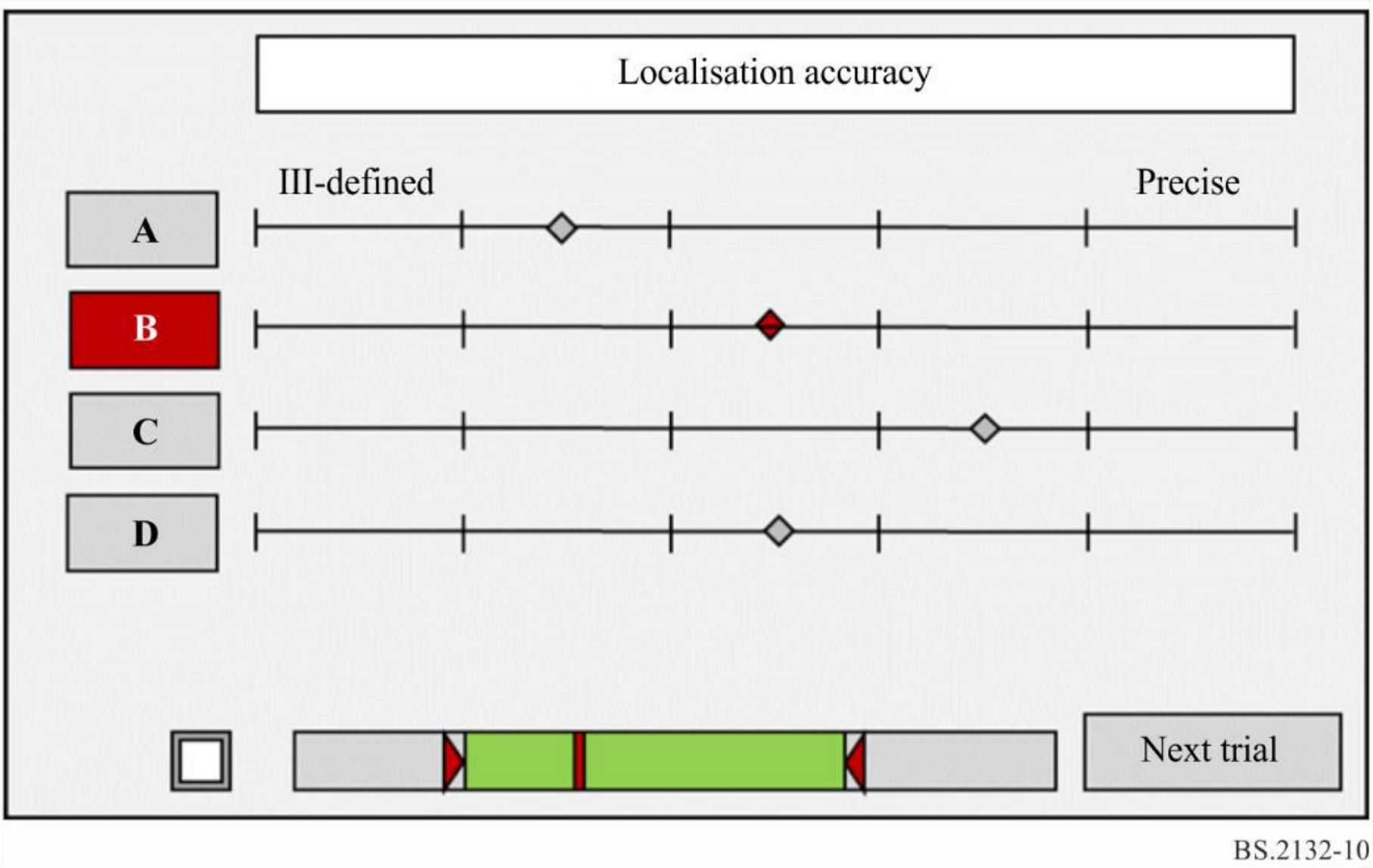
W takim przypadku **nie jest wymagane stosowanie kotwicy (ang. anchor), lub wręcz czasami jej stosowanie jest niepożądane**, na przykład ze względu na charakter badania niemający związku z występowaniem zniekształceń w sygnałach badanych.

Poza tym, przebieg testu jest zbliżony do testu MUSHRA. Także w jego przypadku częste jest przeprowadzanie procedury za pomocą programu komputerowego.

Rekomendacja ITU-R BS.2132

Zazwyczaj pytania zadawane w ramach testów zgodnych z tą rekomendacją mogą dotyczyć:

- Ogólnej subiektywnej jakości
- Oceny subiektywnych parametrów takich jak:
  - głębokość sceny stereofonicznej,
  - Jakość lokalizacji stereofonicznej,
  - wrażenie otoczenia dźwiękiem,
  - jasność dźwięku,
  - stopień przesterowania,



Interfejs graficzny oprogramowania do przeprowadzania testu zgodnego z rekomendacją BS.2132.

źródło: Rekomendacja ITU-R BS.1534.

Rekomendacja ITU-R BS.1285

Podtytuł rekomendacji: Pre-selection methods for the subjective assessment of small impairments in audio systems

Jest to rekomendacja, która pomaga **wybrać ze zbioru wielu systemów te, które wprowadzają największe zniekształcenia.**

Sama ocena systemów przebiega **zgodnie z metodologią** zawartą w rekomendacji **ITU-R BS.1116**, do określania **małych zniekształceń** (test ABC).

Podobnie **analiza** danych ma **przebieg analogiczny do tej rekomendowanej przez dokument ITU-R BS.1116.**

Przeprowadzanie i analiza wyników testów subiektywnych

Rekomendacja ITU-R BS.2126

Podtytuł rekomendacji: Methods for the subjective assessment of sound systems with accompanying picture

Rekomendacja ta definiuje **szereg wymogów na przykład pod względem umiejscowienia i rozmiaru ekranu**, na którym prezentowany jest **materiał wideo**.

Dodatkowo określone są **zasady umiejscawiania monitorów odsłuchowych** względem ekranu i reguły użycia na przykład ekranów, które są transparentne dla fal akustycznych.

W kwestii bardziej szczegółowych zaleceń odnośnie sygnałów audio – rekomendacja odsyła do innych, szczegółowych rekomendacji ITU w tej dziedzinie.

Przykład analizy, test AB porównań parami (lepszy/gorszy)

Test polegał na **wysłuchaniu przez uczestnika testu 20 par próbek**.

W **każdej parze próbek losowo pierwsza, albo druga** z nich byłą próbką **referencyjną**. Druga próbka była próbką zdegradowaną.

Uczestnik miał za zadanie wskazać tę próbkę, która według niego brzmiała lepiej.

Test posiadał **dodatkowe zabezpieczenie służące do testowania wiarygodności odpowiedzi** słuchacza. Był podzielony na **dwie serie**. **Pierwsza seria próbek zawierała 10 porównań** par próbek.

Druga seria zawierała te same pary próbek, jednak w **zmienionej kolejności. Losowo zamienione były także kolejności próbek A oraz B.**

Osoba dokonująca analizy musi stwierdzić, czy słuchacz faktycznie był w stanie usłyszeć różnicę między próbkami i czy faktycznie próbki zdegradowane były przez niego określone jako gorzej brzmiące.

Przeprowadzanie i analiza wyników testów subiektywnych

Przykład analizy, test AB porównań parami (lepszy/gorszy)

R próbka referencyjna Z próbka zdegradowana

seria pierwsza

seria druga

Przeprowadzanie i analiza wyników testów subiektywnych

Przykład analizy, test AB porównań parami (lepszy/gorszy)

Dane tego typu można przeanalizować poprzez sprawdzenie **w których momentach osoba biorąca udział w teście udzieliła odpowiedzi zgodnej ze stanem faktycznym**. Odbywa się to przy założeniu że różnice są słyszalne i próbka zniekształcona brzmi „gorzej” niż próbka referencyjna.

Jeżeli zgodność z takim stanem oznaczymy przez 1, a niezgodność przez zero, wyniki testów można oznaczyć przedstawić w postaci tabelki:

nr próbki	1	2	3	4	5	6	7	8	9	10
zgodność ze wzorcem	1	1	1	0	1	1	1	1	1	1

Odpowiedź powiązana z położeniem próbki referencyjnej została uzyskana w 9 na 10 prób.

Przykład analizy, test AB porównań parami (lepszy/gorszy)

W podobny sposób można przeanalizować odpowiedzi z drugiej serii pytań, tylko w tym przypadku 1. oznacza pokrywanie się odpowiedzi z 2. serii z odpowiedzią z powiązanego pytania z 1. serii.

nr Próbki	1	2	3	4	5	6	7	8	9	10
zgodność odpowiedzi z obu serii	1	1	1	1	1	1	1	1	1	1

Odpowiedź z serii 2. zgodna z odpowiedzią z serii 1. została udzielona 10 razy na 10 prób.

Ciąg tego typu opisany jest tzw. rozkładem Bernoulliego, a analityk chce udowodnić, że prawdopodobieństwa wystąpienia zer i jedynek nie są równe 0,5.

Przykład analizy, test AB porównań parami (lepszy/gorszy)

W przypadku zachodzenia takiej równości w przypadku pierwszej tabelki oznacza to, że próbki **A i B są względem siebie nierozróżnialne**, **natomiast w przypadku analizy powtórzonych odpowiedzi – że nie ma powiązania pomiędzy odpowiedziami z pierwszej i drugiej serii**, czyli **osoba biorąca udział w teście nie słyszała różnic pomiędzy próbkami**

Test ten można przeprowadzić za pomocą testu dwumianowego, który przyjmuje dane z wymienionych tabelek w nieco zmienionej formie – liczby tzw. „sukcesów”, czyli jedynek w ciągu

Przykład analizy, test AB porównań parami (lepszy/gorszy)

Dla odpowiedzi z **1. serii mamy 9 sukcesów na 10 prób**. Przy hipotezie zerowej, że prawdopodobieństwo jedynki równe jest **0,5 p-wartość testu dwumianowego wynosi 0.0214.** Zakładamy, że poziom istotności α wynosi 0,05.

Zatem **odrzucamy hipotezę zerową**, że prawdopodobieństwa wystąpienia zer i jedynek są równe. **Przyjmujemy że są różne. Jako że częstsza jest wartość 1 – przyjmujemy, że różnica jakości pomiędzy próbkami jest słyszalna.**

Podobnie w **przypadku analizy stabilności**, odpowiedź zgodna z odpowiedzią z 1. serii udzielona została **10 razy w ciągu 10 prób**. **Przy hipotezie zerowej identycznej jak w poprzednim teście test dwumianowy zwraca wartość 0,0019, co także dowodzi, że prawdziwa jest hipoteza alternatywna.** Odpowiedzi osoby biorącej udział w teście są stabilne, a zatem – wiargodne.

Obróbka danych – test bazujący na siedmiostopniowej skali dwukierunkowej

Metodologia jest w tym przypadku podobna do analizy przy pomocy testu dwumianowego. Prezentacja próbek odbywa się w analogiczny sposób.

Pierwszą różnicą jest **skala ocen zawierająca się w przedziale od wartości -3** (próbka A jest bardziej zdegradowana) **do wartości +3** (próbka B jest bardziej zdegradowana).

Drugą różnicą, że próbki porównywane z próbką referencyjną mogą wcale nie być zdegradowane lub być zdegradowane w różnym stopniu odpowiadającym bezwzględnym wartościom skali ocen:

- **0** – próbka **referencyjna**,
- **1** – próbka zdegradowana w stopniu **niewielkim**,
- **2** – próbka zdegradowana w stopniu **średnim**,
- **3** – próbka zdegradowana w stopniu **znacznym**.

Obróbka danych – test bazujący na siedmiostopniowej skali dwukierunkowej

Możemy zebrać wyniki przykładowego testu w postaci tabelki:

wzorzec odpowiedzi	-3	2	1	3	1	2	2	-2	0	1
odpowiedzi z 1. serii	-2	2	0	3	2	1	2	-1	0	1
odpowiedzi z 2. serii	-2	3	1	3	2	2	1	-1	0	1

W przypadku takiego testu interesujące może być zbadanie, czy ciąg **różnic pomiędzy wzorcem odpowiedzi a odpowiedziami z 1. serii ma średnią równą zeru. Jeśli tak** by było, oznacza to, że **wzorce** te są ze sobą **zgodne**.

Podobnie można zbadać **zgodność odpowiedzi w 1. i 2. serii** testu.

Hipotezę zerową testu w obydwu przypadkach będzie stanowić **stwierdzenie, że średnia różnic wartości pomiędzy zadanymi ciągami jest równa zeru.**

Obróbka danych – test bazujący na siedmiostopniowej skali dwukierunkowej

Ciągi różnic wyglądają następująco:

różnica między wzorcem a odpowiedziami 1. serii	-1	0	1	0	-1	1	0	-1	0	0
różnice pomiędzy odpowiedziami obu serii	0	-1	-1	0	0	-1	1	0	0	0

Obróbka danych – test bazujący na siedmiostopniowej skali dwukierunkowej

Test t-Studenta dla pierwszego ciągu zwraca **p-wartość równą 0,678**, co oznacza, że **brak jest podstaw dla odrzucenia hipotezy zerowej**. **Zatem wartość średnia analizowanego ciągu, przy wartości poziomu istotności α równej 0,05 jest równa zeru.**

W przypadku **analizy stabilności p-wartość zwrócona przez test t-Studenta wynosi 0,343**. **Zatem w tym przypadku także brak jest podstaw do odrzucenia hipotezy zerowej.**

Należy uznać, że analizowane **odpowiedzi są wiarygodne**, a **wzorzec odpowiedzi wprowadzony przez uczestnika test jest zgodny z faktycznym wzorcem**, według którego przygotowany został test statystyczny.

Obróbka danych – analiza danych z testów podobnych do MUSHRA

Ostatnim przykładem analizy jest **test statystyczny wzorowany na metodologii MUSHRA lub ITU-R BS.2132.**

Prezentowane są podobnie jak w poprzednim przypadku **próbki audio o czterech stopniach zdegradowania – od braku do znacznego.**

Osoba biorąca udział w teście może posłuchać **dowolną ilość razy każą z nich i przypisać jej ocenę od 0 do 10, gdzie 0 oznacza brak przesterowania, a 10 całkowite przesterowanie.**

Nie wie z góry, **która z próbek jest próbką bez zniekształceń**, a która jest zniekształcona w stopniu znacznym

W takim układzie **próbka bez zniekształceń** jest **próbką stanowiącą sygnał referencyjny**, a **próbka zniekształcona w stopniu znacznym** jest sygnałem **kotwicy** (ang. anchor).

Przeprowadzanie i analiza wyników testów subiektywnych

Obróbka danych – analiza danych z testów podobnych do MUSHRA

Wzorzec według których ułożone są próbki

	A	B	C	D
1	brak	mały	średni	znaczny
2	średni	znaczny	mały	brak
3	mały	brak	średni	znaczny
4	średni	brak	znaczny	mały
5	średni	znaczny	brak	mały

...

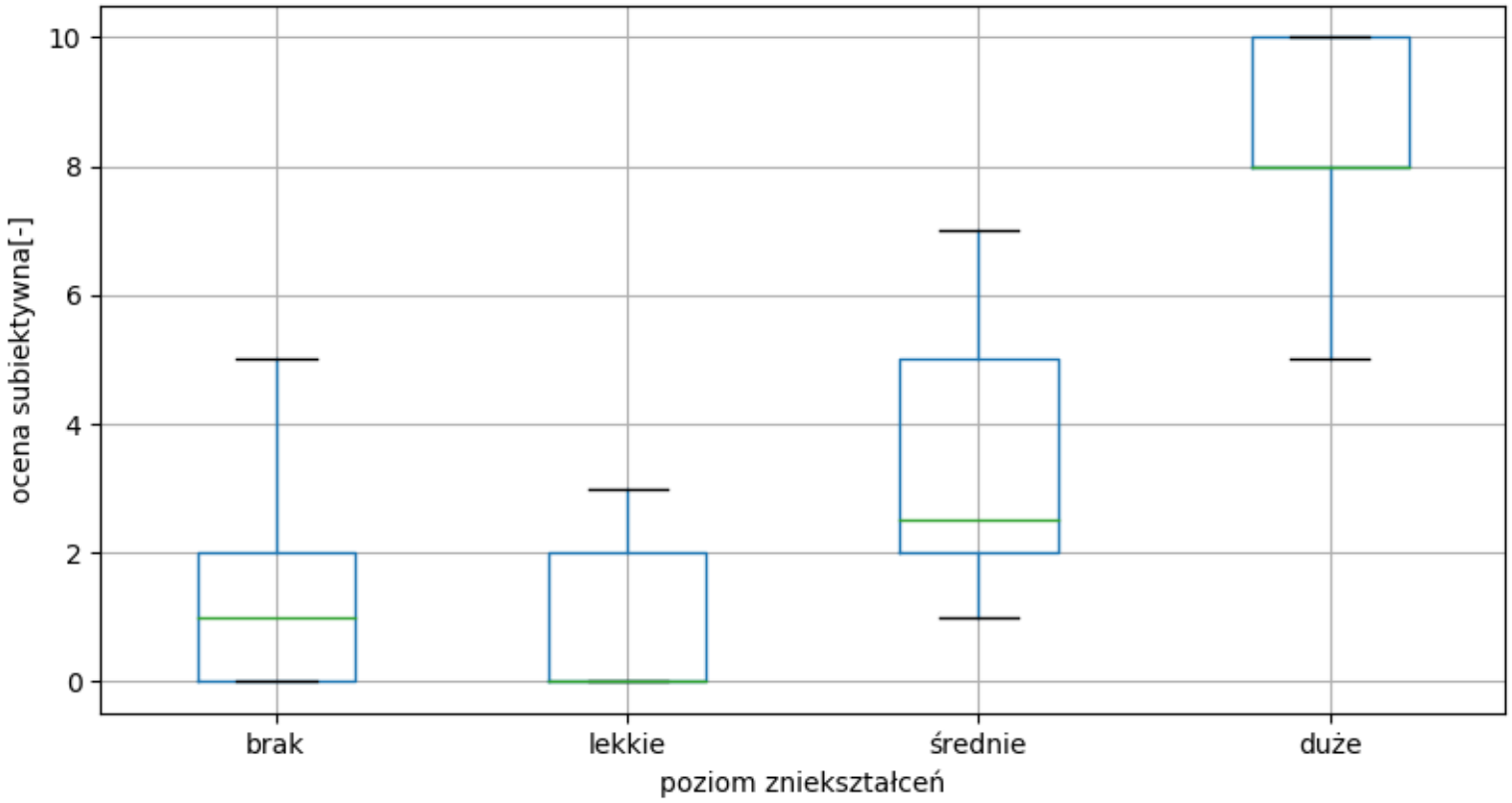
Odpowiedzi osoby biorącej udział w teście

	A	B	C	D
1	0	0	2	7
2	3	10	1	0
3	2	0	5	9
4	4	1	8	2
5	4	10	0	3

...

Przeprowadzanie i analiza wyników testów subiektywnych

Obróbka danych – analiza danych z testów podobnych do MUSHRA



Obróbka danych – analiza danych z testów podobnych do MUSHRA

Dane z testów tego typu bardzo dobrze nadają się do wizualizacji za pomocą **wykresu pudełkowego** – każdy stopień zniekształcenia wiąże się z pojedynczym elementem wykresu.

Analiza polega na **przeprowadzeniu testu statystycznego na różnicę średnich lub median** poszczególnych grup odpowiedzi (dla różnych stopni zniekształcenia).

W przypadku danych zaprezentowanych posłużono się testem **Kruskala-Wallisa**, w przypadku testu **ANOVA** konieczne byłoby uprzednie wykonanie testu **Browna-Forsythe'a** na **równość wariancji** ocen wszystkich 4 grup sygnałów.

Hipotezą zerową testu Kruskala-Wallisa jest brak różnic median poszczególnych grup sygnałów. Test zwrócił p-wartość mniejszą od 10^{-3}, zatem różnice pomiędzy grupami widocznymi na wykresie są istotne statystycznie co najmniej w przypadku jednego porównania.

Przeprowadzanie i analiza wyników testów subiektywnych

Obróbka danych – analiza danych z testów podobnych do MUSHRA

Do ustalenia, które grupy różnią się między sobą wykorzystano test post-hoc Dunn, który zwrócił następującą **macierz p-wartości**:

	brak	**lekkie**	**średnie**	**znaczne**
Brak		0,579	0,009	$< 10^{-3}$
Lekkie	0,579		0,002	$< 10^{-3}$
Średnie	0,009	0,002		$< 10^{-3}$
znaczne	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	

Istotne są zatem wszystkie różnice poza poziomem braku i lekkich zniekształceń. Grupa referencyjna otrzymała wartość najniższą, a kotwica najwyższą, zatem można stwierdzić, że **otrzymane wyniki są wiarygodne.**

Zobiektywizowane miary jakości

Wadą testów odsłuchowych jest fakt, iż są one **czasochłonne i drogie**. Z tego względu poszukiwano **możliwych sposobów na przewidzenie ich wyniku**, **za pomocą obiektywnych algorytmów**. Istnieje kilka algorytmów, które w sposób zobiektywizowany obliczają estymowaną wartość oceny subiektywnej dla zadanych sygnałów audio:

- **PSQM (Perceptual Speech Quality Measurement)** – najstarszy test, służący do badania jakości mowy w systemach telekomunikacyjnych. Wynikiem jego działania jest wynik w skali MOS. Algorytm bazuje na szeregu przekształceń wykorzystujących między innymi skalę barkową i model psychoakustyczny. **Zakres zwracanych wartości MOS – od 1,0 (najgorsza jakość) do 5,0 (najlepsza jakość).**
- **PESQ (Perceptual Evaluation of Speech Quality)** – test nowszy od PSQM, **między innymi uwzględnia pakietową transmisję danych**. Uwzględnia między innymi takie **zjawiska takie jak jitter**. Zaleca się wykorzystywanie tej miary w miejsce PSQM. **Podobnie jak PSQM PESQ zwraca wartość MOS, z tym że w przedziale od 1,0 do 4,5.**

Zobiektywizowane miary jakości

- **PEAQ (Perceptual Evaluation of Audio Quality)** – test pozwalający na **ocenę sygnałów szerokopasmowych takich jak sygnały muzyczne**. Bazuje on na **modelu psychoakustycznym wykorzystującym sztuczną sieć neuronową**. W zależności od modelu sztucznej sieci neuronowej istnieją **dwie wersje testu**, prostsza – PEAQ Basic i PEAQ Advanced. Wersja Basic jest mniej dokładna, ale mniej złożona obliczeniowo, wersja Advanced daje dokładniejsze wyniki. **Wynik algorytmu stanowi zespół parametrów opisujący różnicę jakości** pomiędzy sygnałem zniekształconym a referencyjnym (algorytm dokonuje porównania), **czy też szereg miar opisujący występujące w sygnale wejściowym zniekształcenia**.

Dziękuję za uwagę!