Metody oceny jakości dźwięku

Testy subiektywne, PESQ, PEAQ

Grzegorz Szwoch, Katedra Systemów Multimedialnych, Politechnika Gdańska

Wprowadzenie

- Problem: ocena jakości dźwięku – urządzenia (np. zestawy głośnikowe), pomieszczenia, algorytmy
- Metody obiektywne - np. pomiar SNR czy THD+N - nie dają pełnych informacji o jakości sygnału.
- Ważne jest to, jaka jest jakość dźwięku odbieranego subiektywnie przez słuchacza.
- Stosuje się testy subiektywne.
- Grupa słuchaczy odsłuchuje testowy zestaw sygnałów i ocenia każdy z nich według podanej skali.
- Jest to ocena subiektywna, ponieważ opiera się na indywidualnym wrażeniu odnoszonym przez słuchacza.

Wprowadzenie

Przykładowe zastosowania testów subiektywnych – ocena:

- jakości sprzętu audio (kolumny głośnikowe, odtwarzacze)
- jakości dźwięku w pomieszczeniu
- jakości transmisji mowy lub muzyki
- jakości kodowania mowy/muzyki (np. kompresja stratna)
- wyników działania algorytmu (np. redukcja szumu, rekonstrukcja starych nagrań)

W tych (i podobnych) przypadkach testy subiektywne są jedyną możliwością uzyskania wiarygodnej oceny jakości sygnału dźwiękowego.

Etapy testu subiektywnego

- Przygotowanie testu
  - procedura testowa
  - zestaw sygnałów testowych
- Zgromadzenie grupy słuchaczy, weryfikacja
- Trening – zapoznanie z procedurą testową
- Testy – każdy słuchacz dokonuje oceny
- Analiza wyników (zwykle statystyczna)
  - weryfikacja słuchaczy (po teście)
- Raport

Wiarygodność wyników testu

Czynniki wpływające na wiarygodność wyników testu:

- odpowiednio dobrana grupa słuchaczy
- odpowiedni dobór sygnałów testowych (muszą być reprezentatywne dla badanego przypadku)
- odpowiedni wybór procedury testowej: czas trwania, metoda prezentacji, sposób oceniania
- warunki w pomieszczeniu testowym (brak zakłóceń)
- jakość sprzętu (głośniki, słuchawki)
- usytuowanie głośników i słuchacza w pomieszczeniu
- poziom głośności prezentowanych sygnałów
- poprawna analiza statystyczna i weryfikacja wyników

Dobór grupy słuchaczy

Eksperci czy zwykli słuchacze?

Zwykły słuchacz:

- reprezentuje „statystycznego użytkownika” oraz „docelowego konsumenta”
- można wykonywać testy preferencji: „A czy B”
- nie ma doświadczenia w ocenianiu jakości, nie wie jak przydzielać oceny („4,4 czy 4,5?”)
- w większości przypadków wyniki uzyskane od takich słuchaczy byłyby odrzucone na etapie weryfikacji po teście – czyli nie mamy wyników

Dobór grupy słuchaczy

Eksperci czy zwykli słuchacze?

Ekspert:

- osoba wyszkolona do „wychwytywania” zniekształceń i różnic między sygnałami
- ma doświadczenie – potrafi znaleźć „punkt odniesienia” przy nadawaniu oceny
- prawidłowy słuch (kontrola audiometryczna)
- jego oceny są wiarygodne

Zaleca się przeprowadzanie testów z grupą ekspertów.

Najczęściej wystarcza uzyskać wyniki od ok. 20 ekspertów.

Weryfikacja słuchaczy

Weryfikacja – eliminowanie niewiarygodnych słuchaczy

Przed testem – na podstawie:

- wyników wstępnych testów (np. podczas treningu)
  - wychwycenie różnic zauważalnych dla ekspertów, a nie zauważalnych dla zwykłego słuchacza
- badań audiometrycznych
- doświadczenia w testach subiektywnych

Po teście – na podstawie analizy wyników:

- wyniki znacząco różne od średniej
- niewiarygodne wyniki, np. niepowtarzalne, lub sygnały zniekształcone oceniane nie niżej niż oryginalne

Skala ocen

Skala ocen musi być z góry ustalona i taka sama dla wszystkich słuchaczy. Często stosuje się skalę od 1 (najgorsza jakość) do 5 (najlepsza jakość), z krokiem co 0,1 (a więc 41 stopni).

Ocena	Jakość	Zniekształcenia
5	Doskonała	Niesłyszalne
4	Dobra	Słyszalne, nie dokuczliwe
3	Średnia	Lekko dokuczliwe
2	Słaba	Dokuczliwe
1	Zła	Bardzo dokuczliwe

Skala ocen

Skala ocen względnych, stosowana w celu porównania jakości jednego sygnału względem innego (uwaga: wyniki nie są porównywalne z poprzednią metodą). Krok 0,1 (61 st.).

Ocena	Porównanie jakości
3	Znacznie lepsza
2	Lepsza
1	Nieco lepsza
0	Taka sama
–1	Nieco gorsza
–2	Gorsza
–3	Znacznie gorsza

Faza treningu

Przed fazą oceny wykonuje się trening:

- Objaśnienie procedury testowej (co ma być oceniane, jak będą prezentowane sygnały, w jaki sposób mają być oceniane, itp.) – najlepiej bezpośrednio (ustnie), może być też pisemna lub odtwarzana instrukcja.
- Demonstracja procedury testowej – przykładowe sygnały (nie wykorzystane później w teście, ale prezentowane w ten sam sposób). Można wskazać które sygnały są oryginalne, a które zniekształcone. Można też wykorzystać fazę treningu do weryfikacji słuchacza.

Konieczna jest przerwa między fazą treningu a zasadniczym testem.

Procedura testowa

Test może być przeprowadzany w różny sposób.

- Wszyscy słuchacze oceniają jednocześnie
  - o ile warunki i sposób przeprowadzania testu na to pozwalają
  - uwaga na warunki – np. miejsce w sali może wpływać na ocenę!
- Każdy słuchacz ocenia osobno. Jedna ciągła sesja na słuchacza.
  - znacznie dłuższy czas testów
  - warunki muszą być jednakowe
  - nie ma problemów z dostępnością słuchaczy

Procedura testowa

- Czas trwania testu nie powinien być zbyt długi (zmęczenie słuchacza).
- Maksymalny czas trwania sesji: 20 – 30 minut (bez przerwy)
- Maksymalnie 10 – 15 ocenianych sygnałów w sesji.
- Jeżeli potrzebna jest więcej niż jedna sesja, przerwa między sesjami musi trwać przynajmniej tyle co jedna sesja.

Prezentacja sekwencji testowych

Sygnały testowe mogą być prezentowane w sposób:

- automatyczny (narzucony z góry)
  - każda oceniana sekwencja powinna być powtórzona
  - konieczne przerwy na dokonanie i wpisanie oceny
  - metoda stosowana zwykle przy grupie słuchaczy oceniającej sygnały w tym samym czasie
- interaktywny
  - słuchacz sam decyduje o liczbie powtórzeń
  - najczęściej „badamy” tą metodą jednego słuchacza w danej chwili
  - dłuższy test, ale wyniki mogą być bardziej wiarygodne

Prezentacja sekwencji testowych

Sposoby prezentacji sygnałów testowych i ich oceny:

- pojedynczy sygnał – ocena jego jakości (problem punktu odniesienia)
- porównanie parami – ocena różnicy między dwoma sygnałami
- porównanie sygnału ocenianego z referencyjnym
- porównania z wykorzystaniem ukrytego sygnału referencyjnego (np. test ABC)
- porównanie zestawu ocenianych sygnałów z sygnałem referencyjnym (np. test MUSHRA).

Przygotowanie sekwencji testowych

- Sygnał referencyjny (*reference*) – sygnał „oryginalny” (nie zniekształcony) – do porównania z sygnałem ocenianym.
- Sygnały testowe muszą być tak dobrane, aby możliwe było uzyskanie żądanej informacji.
- Nie można dobierać sygnałów testowych w sposób „tendencyjny”, tak aby uzyskać pożądane wyniki.
- Pojedynczy sygnał nie może trwać dłużej niż 15-20 sekund (ale może być znacznie krótszy).
- Sygnały (zwłaszcza muzyka i mowa) nie mogą się gwałtownie urywać (musi być np. wyciszenie).

Przygotowanie sekwencji testowych

- Między sygnałem referencyjnym a ocenianym: przerwa ok. 0,5 – 1 sekundy.
- Między powtórzeniami sekwencji „referencyjny – oceniane”: dłuższa przerwa, ok. 1 – 1,5 sekundy.
- Jeżeli prezentowany jest ten sam sygnał z różnymi zniekształceniami – nie pod rząd.
- Kolejność sekwencji testowych powinna być losowa, różna dla kolejnych słuchaczy (eliminacja wpływu zmęczenia na wyniki testu).
- „Treść” sygnałów powinna być neutralna, tak aby preferencje słuchacza (np gatunek muzyczny) nie wpływały na ocenę.

Oceniane atrybuty

- Ocenie podlegają parametry sygnału – atrybuty.
- Podstawowy atrybut, zawsze oceniany:
  - jakość sygnału (*basic audio quality*), lub
  - zniekształcenia (*distortions*).
- Inne atrybuty mogą być definiowane w zależności od testu. Np. test pomieszczenia – oceniane atrybuty mogą dotyczyć: zrozumiałości mowy, selektywności sygnału, sceny stereofonicznej, zawartości pogłosu, itp.
- Nie należy przesadzać z liczbą atrybutów – utrudnia to słuchaczom ocenę.

Testy oceny bezwzględnej

Testy oceny bezwzględnej:

- prezentujemy sygnały testowe
- słuchacz musi ocenić jakość sygnału
- słuchacz musi sam znaleźć „punkt odniesienia”
- np. ocena zestawów głośnikowych, jakości dźwięku w pomieszczeniu
- udział ekspertów bezwzględnie wymagany!
- nie uzyskamy dobrych wyników od zwykłych słuchaczy – nie potrafią oni przydzielać ocen nie mając podanego sygnału odniesienia

Testy oceny bezwzględnej i względnej

Ocena względna – testy zniekształceń:

- oceniana jest różnica między sygnałem referencyjnym a ocenianym – zniekształcenie
- prostsze do oceny – sygnał odniesienia jest podany wprost
- nadaje się do testów porównawczych „A czy B?”
- np. ocena działania algorytmy poprawy jakości dźwięku „przed i po”
- można stosować dla zwykłych słuchaczy („kolegów studentów”), ale trzeba zwiększyć liczbę słuchaczy i dokładnie sprawdzić ich wiarygodność

Ocena zniekształceń - test AB

- Zniekształcenie (*impairment*) – różnica między sygnałem referencyjnym a ocenianym.
- Jeżeli zniekształcenia są duże, wystarczy porównywać parami sygnał oceniany z referencyjnym i przydzielać oceny na podstawie porównania.
- Test AB – prezentujemy słuchaczom sekwencje AB (z powtórzeniami):
  - A – sygnał referencyjny
  - B – sygnał oceniany względem A
- Ta metoda zawiedzie w przypadku gdy różnica między A i B jest mała

Ocena zniekształceń – test ABC

- Metoda „podwójnie ślepej próby trzech pobudzeń z ukrytym sygnałem referencyjnym” (test ABC)
- Może być stosowana również w przypadku małych zniekształceń (niewielkich różnic).
- Prezentujemy słuchaczom trójki sygnałów:
  - A: zawsze sygnał referencyjny (nie oceniany)
  - B: pierwszy sygnał oceniany względem A
  - C: drugi sygnał oceniany względem A
- Słuchacz wpisuje oceny (względne) dla B i C.

Ocena zniekształceń – test ABC

- Jeden z sygnałów – losowo B lub C – jest sygnałem referencyjnym, drugi to sygnał podlegający ocenie.
- Słuchacz nie wie który sygnał jest referencyjny, a który ma być oceniany!
- Spodziewamy się, że sygnał referencyjny uzyska najlepszą ocenę (brak różnic).
- Oceny przydzielane sygnałom referencyjnym są wykorzystywane do weryfikacji słuchaczy po teście.
- Jeżeli słuchacz uporczywie „słyszy” różnice między referencyjnym B lub C a referencyjnym A, to znaczy że nie jest wiarygodny i należy go wykluczyć.

Ocena zniekształceń – test MUSHRA

- Metoda ABC zawodzi w przypadku silnie zniekształconych sygnałów (np. duża kompresja) – wyniki są skumulowane w dolnej części skali.
- W tym przypadku stosuje się metodę MUSHRA.
- Słuchacz odsłuchuje grupy sygnałów – może porównywać między sobą sygnały o różnym stopniu zniekształcenia.
- Metoda stosowana dla pojedynczych słuchaczy.

Ocena zniekształceń – test MUSHRA

Grupa sygnałów ocenianych przez słuchacza zawiera:

- sygnał referencyjny (ukryty)
- oceniane sygnały (różne stopnie zniekształceń),
- punkty zaczepienia (*anchors*):
  - sztucznie zniekształcony sygnał referencyjny (np. ograniczenie pasma – filtracja dolnoprzepustowa, dodany szum, itp.);
  - co najmniej jeden taki sygnał (zwykle 2 – 3)

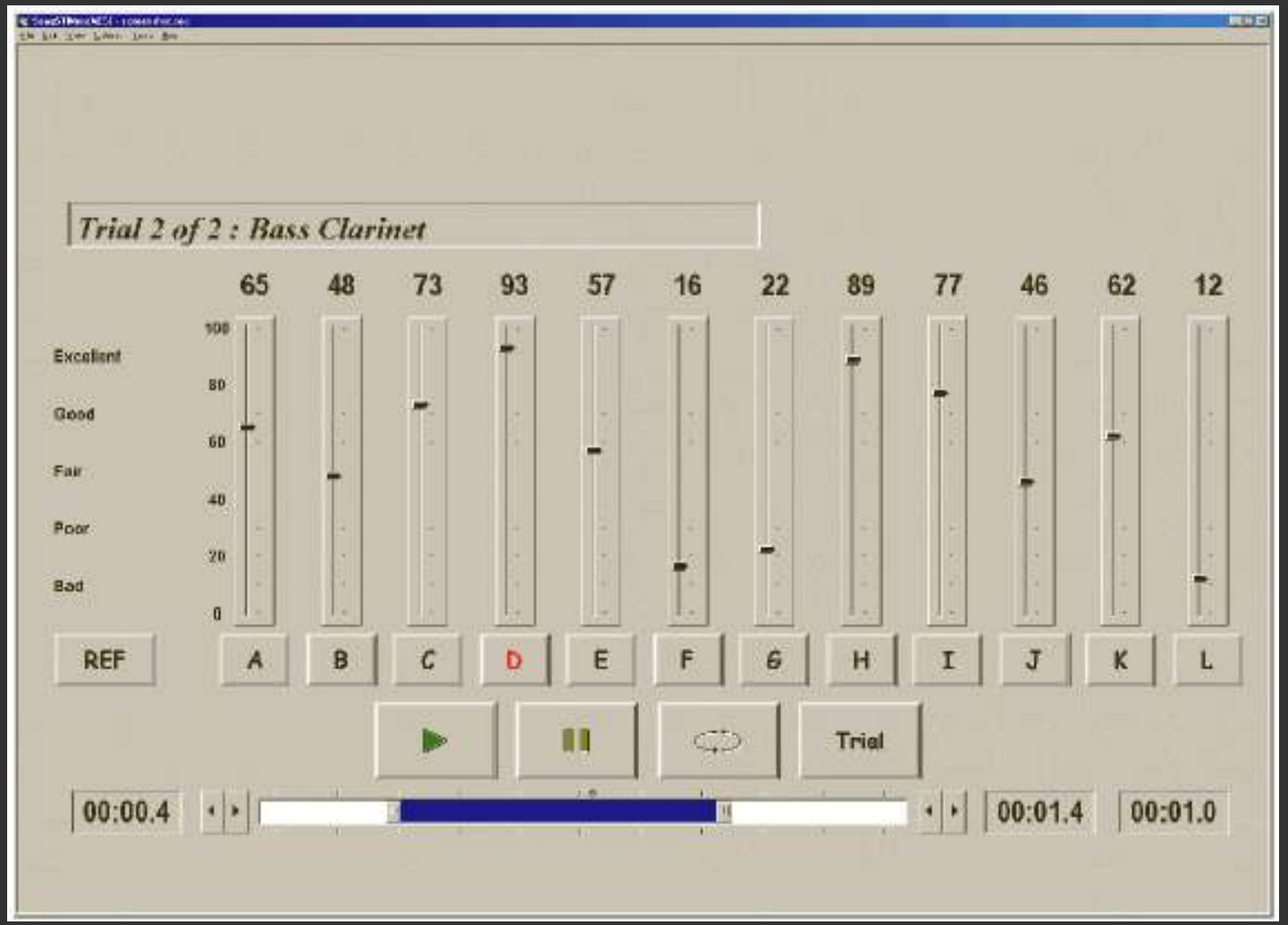
Słuchacz ocenia każdy sygnał.

Ocena zniekształceń – test MUSHRA

- W teście MUSHRA słuchacz ocenia każdy sygnał w skali od 0 (najgorsza jakość) do 100 (najlepsza jakość), z krokiem 1.
- Słuchacz powinien mieć możliwość odtworzenia każdego sygnału w ramach grupy dowolną liczbę razy.
- Sekwencja grup sygnałów oraz kolejność sygnałów w grupie powinny być losowe (nie powtarzać się dla kolejnych słuchaczy).
- Grupa powinna zawierać maksymalnie 15 sygnałów, wliczając referencyjny i punkty zaczepienia.

Ocena silnie zniekształconych sygnałów (cd.)

Przykładowy panel służący do oceny sygnałów. Słuchacz nadaje oceny za pomocą suwaków.



Przykład przygotowania testu MUSHRA

- Chcemy ocenić wyniki działania naszego algorytmu poprawy jakości dźwięku z głośnika w laptopach.
- Zbiór testowy – nagrania różnego typu (np. muzyka klasyczna, akustyczna, rockowa, popowa, jazzowa, elektroniczna, mowa).
- Kolejność nagrań – losowana dla każdego słuchacza.
- Sygnały referencyjne – muzyka z płyty.
- Punkty zaczepienia: przesterowujemy, obcinamy pasmo, dodajemy szum.
- Sygnały oceniane – przetworzone przez nasz algorytm, np. różne stopnie poprawy dźwięku.

Ocena testu MUSHRA

- Oceny uzyskane dla rzeczywistych sygnałów testowych są poddawane analizie.
- Przy porównywaniu zniekształceń, sygnał referencyjny powinien uzyskać ocenę 100.
- Sygnały „zaczepienia” (pogorszone) powinny uzyskać najniższe oceny.
- Uzyskane wyniki pozwalają zweryfikować słuchaczy. Jeżeli nie potrafi on wskazać sygnału referencyjnego lub punktów zaczepienia – nie jest wiarygodny, jego oceny należy odrzucić.

Testy subiektywne kodowanej mowy i muzyki

Testy kodowanej mowy

- Oceniana jest jakość sygnału, porównanie z wzorcowymi zniekształceniami.
- Ocena jakości przy użyciu skali MOS - *Mean Opinion Score*: od 1 (zła) do 5 (doskonała), z krokiem 0,1.

Testy kodowanej muzyki

- Porównanie dźwięku zakodowanego z oryginalnym. Ocena zniekształceń w skali od 1 (bardzo dokuczliwe) do 5 (niesłyszalne), z krokiem 0,1.
- Skala SDG (*Subjective Difference Grade*): różnica ocen dla sygnału badanego i referencyjnego, od –4 do 0

Analiza wyników testów

Wyniki analizuje się za pomocą testów statystycznych.

Podstawowe parametry:

- wartość średnia z ocen i wariancja
- przedział ufności (w nim znajduje się większość wyników)
- poziom istotności (zwykle 0,05 – oznacza, że z prawdopodobieństwem 0,95 możemy powiedzieć, że wynik jest zawarty w przedziale ufności).

Do weryfikacji hipotez (np. czy jakość przetworzonego sygnału jest istotnie gorsza od sygnału referencyjnego) stosuje się odpowiednie testy statystyczne, np. test t-Studenta lub analizę wariancji ANOVA.

Raport z testów subiektywnych

Preferowana zawartość raportu:

- cel przeprowadzenia testu
- charakterystyka słuchaczy i sygnałów testowych
- opis warunków akustycznych podczas testu
- dokładny opis procedury testowej
- opis sposobu analizy wyników
- końcowy wynik (w formie wykresów i tabel)
- dokładniejszy opis wyników (jeżeli chcemy zamieścić szczegółowe analizy, powinny być one w załączniku)
- wnioski wynikające z uzyskanych wyników

Problemy związane z testami subiektywnymi

Badanie jakości sygnałów za pomocą testów subiektywnych napotyka na szereg trudności:

- konieczność zgromadzenia grupy ekspertów
- długi czas badania (i koszt)
- zmęczenie słuchaczy i czynniki indywidualne wpływają na wyniki
- konieczność analizy statystycznej wyników w celu uzyskania porównywalnych ocen
- problem powtarzalności wyników
  - dla różnych grup słuchaczy
  - dla tej samej grupy słuchaczy

Obiektywizacja testów subiektywnych

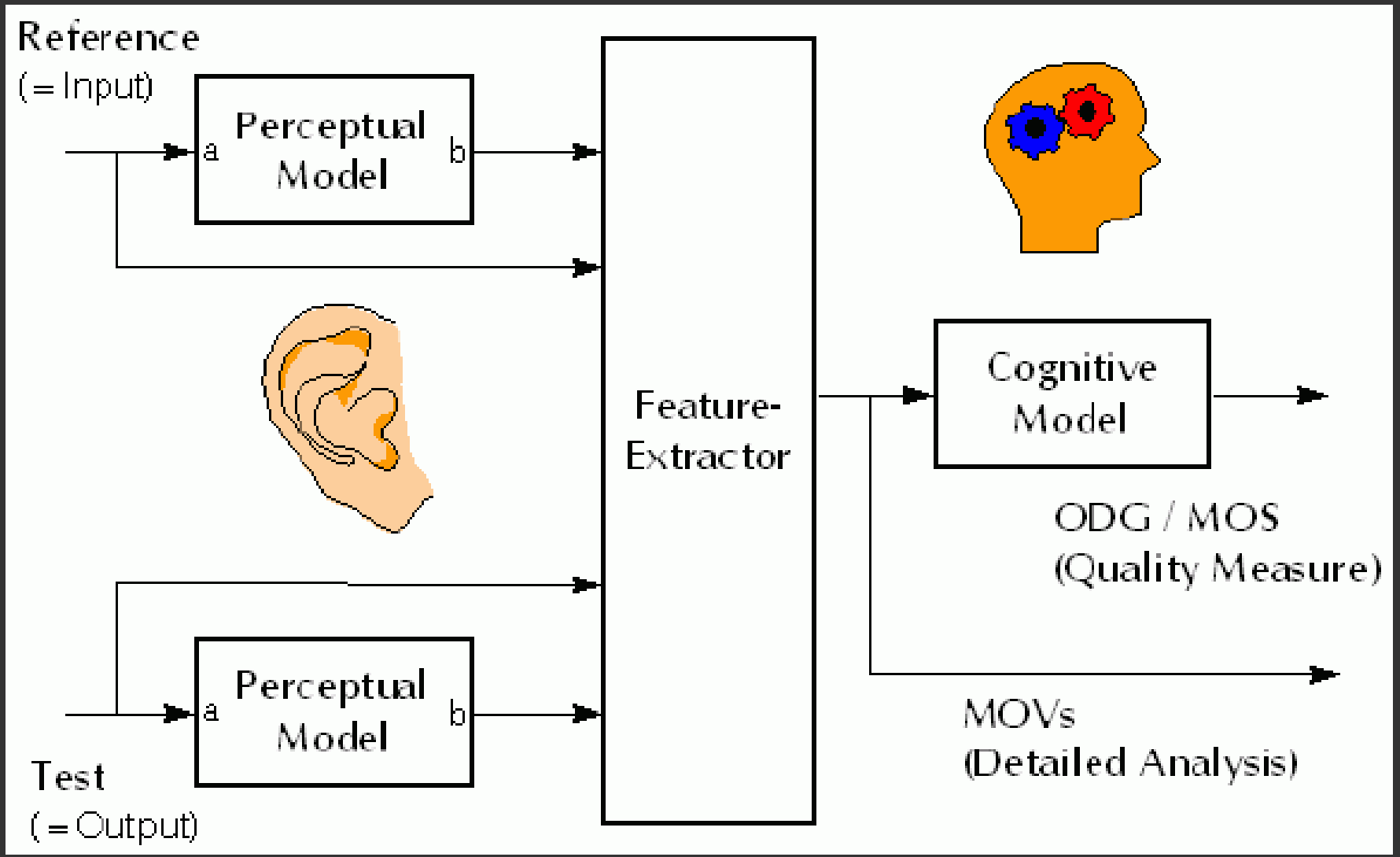
- Obiektywizacja polega na zastąpienie grupy słuchaczy przez program komputerowy: „test zobiektywizowany”.
- Algorytm perceptualny ocenia jakość sygnału w taki sposób, w jaki robi to człowiek.
- Założenie: wysoka zgodność (korelacja) wyników testu „komputerowego” z wynikami testów odsłuchowych.
- Zalety:
  - skrócenie czasu testów
  - wyeliminowanie konieczności pracy z grupą słuchaczy
  - powtarzalność wyników
- Przykład programu: Opticom Opera

Obiektywizacja testów subiektywnych

Etapy algorytmu:

- przetwarzanie wstępne: wyrównanie czasowe, normalizacja amplitudy
- model perceptualny: przetwarzanie przez model słyszenia (pasma krytyczne, maskowanie, itp.)
- ekstrakcja cech: obliczenie istotnych parametrów sygnału
- model kognitywny: algorytm sztucznej inteligencji (np. sieć neuronowa), który przekłada wartości parametrów na ocenę słuchacza

Algorytm perceptualny



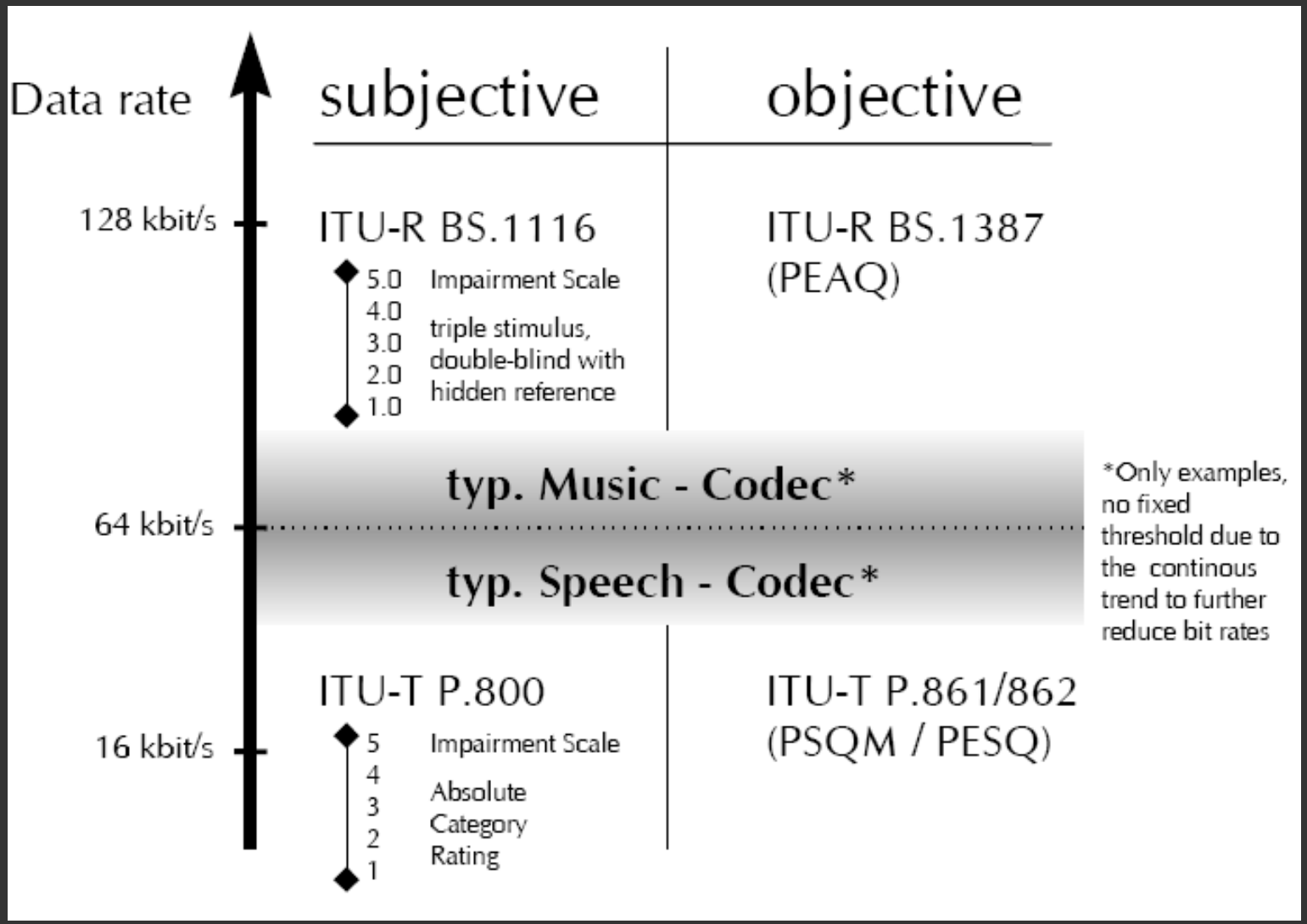
Komputerowe testy subiektywne

Najważniejsze testy używane do oceny jakości dźwięku:

- PSQM – *Perceptual Speech Quality Measurement* – starszy test do badania jakości sygnału mowy w analogowych systemach telekomunikacyjnych
- PESQ – *Perceptual Evaluation of Speech Quality* – nowszy test do badania jakości sygnału mowy, uwzględnia pakietową transmisję danych
- PEAQ – *Perceptual Evaluation of Audio Quality* – badanie jakości sygnałów szerokopasmowych (muzycznych)

Testy odsłuchowe i komputerowe

Zalecenia: kiedy i jak stosować każdy z testów



PSQM

PSQM - ocena jakości sygnału w systemach telekomunikacyjnych. Metoda nie sprawdza się jednak we współczesnych zastosowaniach, np. VoIP (zmienne opóźnienia pakietów). Struktura algorytmu:

- transformacja sygnału referencyjnego i badanego do skali barkowej
- filtracja uwzględniająca charakterystyki słuchawek i mikrofonów
- symulacja zakłóceń (szum Hotha)
- porównanie sygnałów w modelu psychoakustycznym
- obliczenie miar jakości sygnału (skala MOS)

PSQM

Wyniki testu PSQM

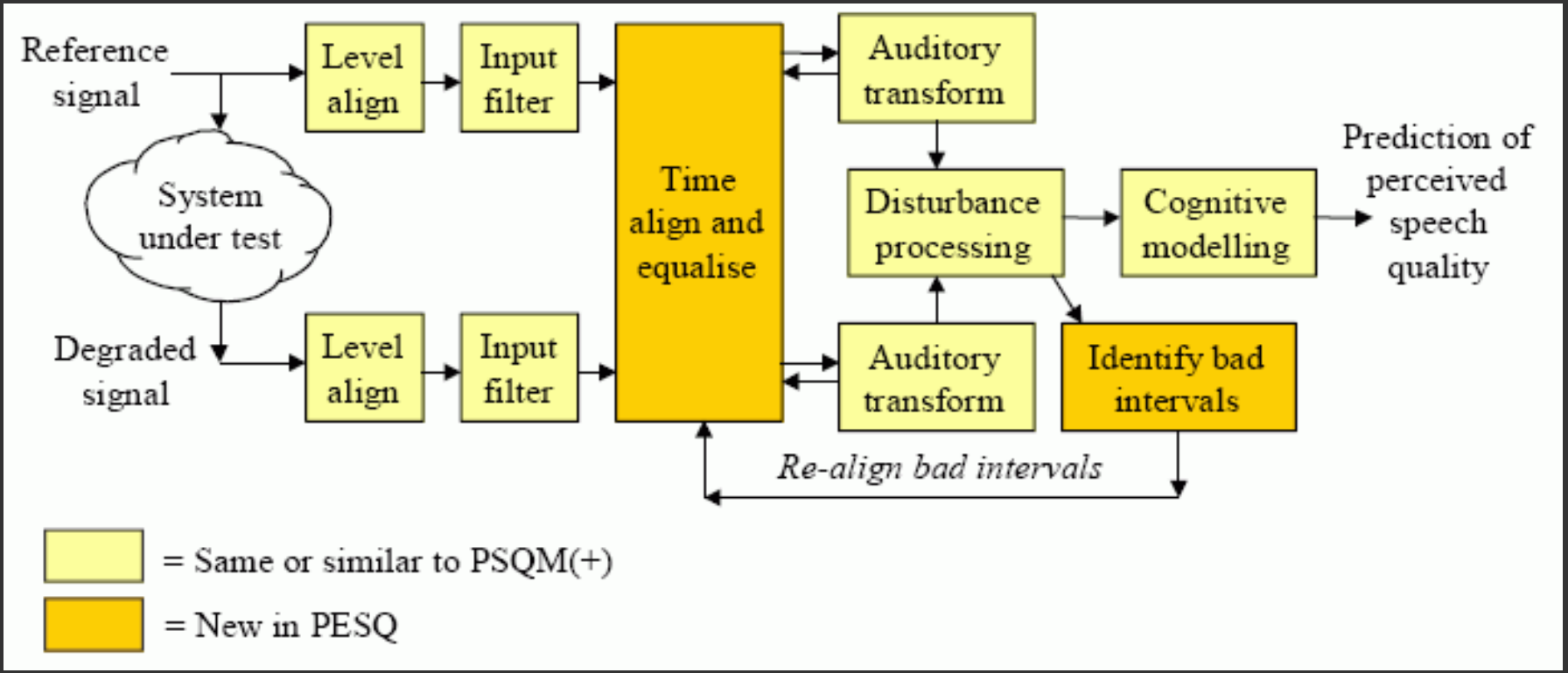
- Miara PSQM - ocena pogorszenia jakości dźwięku na skutek kodowania i transmisji sygnału
- PSQM-Wx - miara PSQM z uwzględnieniem fragmentów ciszy (wg normy), ze współczynnikiem wagowym 0,x
- OMOS - wynik PSQM przeliczony na skalę testów subiektywnych MOS, od 1.0 (najgorsza jakość) do 5.0 (najlepsza jakość)

Wynik OMOS odpowiada wynikowi testów odsłuchowych, wyrażonemu za pomocą skali MOS.

PESQ

Test PESQ jest rozszerzeniem testu PSQM. Uwzględnia zmienne opóźnienie transmisji (*jitter*), np. w sieciach VoIP. Zaleca się stosowanie testu PESQ zamiast PSQM.

Dodano algorytm wyrównywania czasowego sygnałów.



PESQ

- Wynik testu PESQ: miara MOS w skali od 1.0 (najgorsza jakość) do 4.5 (najlepsza jakość).
- Najlepsza jakość w skali MOS sięga wartości 5.0, ale taki wynik nie jest nigdy osiągany w analizie statystycznej wyników testów subiektywnych.

PEAQ

- PEAQ - ocena jakości dźwięku szerokopasmowego (np. muzyka).
- Istnieją dwie wersje testu.
  - *PEAQ Basic* - uproszczona wersja, mniej dokładny test, ale analiza może być przeprowadzana w czasie rzeczywistym.
  - *PEAQ Advanced* - wymaga bardziej skomplikowanych obliczeń i dłuższego czasu analizy, ale daje dokładniejsze wyniki
- Struktura obu wersji jest podobna. Wersje różnią się złożonością modelu psychoakustycznego, opartego na sztucznej sieci neuronowej.

PEAQ

PEAQ porównuje sygnał referencyjny (oryginalny)
z sygnałem po zakodowaniu i odkodowaniu:

- obliczenie FFT i skalowanie sygnałów
- modelowanie wpływu ucha zewnętrznego i środkowego (za pomocą filtrów)
- transformacja do skali barkowej (pasma krytyczne)
- uwzględnienie zjawiska maskowania
- porównanie sygnałów w modelu perceptualnym
- ekstrakcja cech sygnałów
- obliczenie wyników testu

PEAQ

- Na podstawie porównania sygnałów uzyskuje się wektor cech – MOV (*Model Output Variable*). Zmienne MOV opisują poszczególne parametry, które mają wpływ na jakość dźwięku (wielkość zniekształceń, modulacja, itp.).
- ODG (*Objective Difference Grade*) – różnica jakości pomiędzy dźwiękiem badanym a referencyjnym (odpowiednik SDG w testach odsłuchowych), skala od –4.0 (zniekształcenia bardzo dokuczliwe) do 0.0 (zniekształcenia niesłyszalne).
- DI (Distortion Index) – inna miara zniekształceń

PEAQ

Zmienne wektora MOV dla obu wersji testu PEAQ

Basic

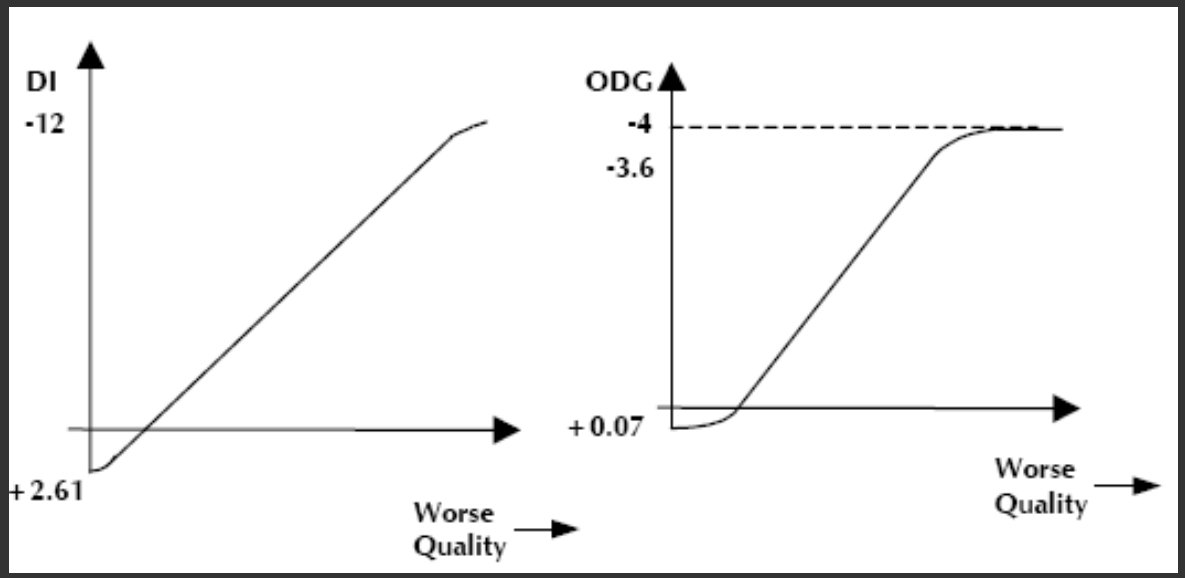
Model Output Variable (MOV)	Interpretation
$WinModDiff1_B$	Changes in modulation (related to roughness)
$AvgModDiff1_B$	
$AvgModDiff2_B$	
$RmsNoiseLoud_B$	Loudness of the distortion
$BandwidthRef_B$	Linear distortions (frequency response etc.)
$BandwidthTest_B$	
$RelDistFrames_B$	Frequency of audible distortions
Total NMR_B	Noise-to-mask ratio
$MFPD_B$	Detection probability
ADB_B	
EHS_B	Harmonic structure of the error

Advanced

Model Output Variable (MOV)	Interpretation
$RmsNoiseLoudAsym_A$	Loudness of the distortion
$RmsModDiff_A$	Changes in modulation (related to roughness)
$AvgLinDist_A$	Linear distortions (frequency response etc.)
Segmental NMR_B	Noise-to-mask ratio
EHS_B	Harmonic structure of the error

PEAQ

Miary ODG i DI są powiązane ze sobą, ale nie można porównywać wartości ODG i DI (zwłaszcza pochodzących z różnych pomiarów) ze sobą.



Zalecenia:

- ODG > –3.6: stosować ODG
- ODG < –3.6: stosować DI

Bibliografia

Normy ITU (www.itu.int) dotyczące testów subiektywnych:

- ITU-R BS.1283-1: wykaz standardów dot. testów subiektywnych
- ITU-R BS.1284: ogólne metody subiektywnej oceny jakości dźwięku
- ITU-R BS.1116: ocena małych zniekształceń dźwięku (test ABC)
- ITU-R BS.1534: ocena jakości w systemach kodowania (MUSHRA)
- ITU-R BS.1285: subiektywne testy preselekcyjne
- ITU-R BS.1286: ocena jakości dźwięku z towarzyszącym obrazem
- ITU-T P.800: subiektywna ocena jakości transmitowanego sygnału
- ITU-T P.861: test PSQM
- ITU-T P.862: test PESQ
- ITU-R BS.1387: test PEAQ