FBMF Technical Report

기술보고서

FBMF-TR-020 제정일: 2025. 12. 05.

ITU-R BT.500 주관적 영상 품질 평가 기술 분석

Analysis of Subjective Video Quality Assessment Technologies Based on ITU-R BT.500

FBMF

미래방송미디어표준포럼

Future Broadcast and Media Standards Forum

표준초안 검토 위원회 UHD 융합기술분과위원회

표준안 심의 위원회 운영위원회

	성명	소 속	직위	위원회 및 직위
기술보고서(과제) 제안	조숙희	ETRI	책임	UHD 융합기술분과위원회 의장
기술보고서 초안 에디터	홍순기	SBS	책임	UHD 융합기술분과위원회 간사
	주재환	SBS	책임	UHD 융합기술분과위원회 위원
	정진우	KETI	책임	UHD 융합기술분과위원회 위원
	고민수	KETI	책임	UHD 융합기술분과위원회 위원
	백형석	LG 전자	책임	UHD 융합기술분과위원회 위원
	오혜주	ETRI	책임	UHD 융합기술분과위원회 위원
사무국 담당	함상진	KBS	책임	운영위원회 사무총장

발행인 : 미래방송미디어표준포럼 의장

발행처 : 미래방송미디어표준포럼

06130, 서울특별시 강남구 테헤란로 7 길 22 신관 1108 호

Tel : 02-568-3556, Fax : 02-568-3557

발행일 : 2025.12.05.

서 문

1 기술보고서의 목적

이 기술보고서는 국제전기통신연합 무선통신부문(ITU-R)에서 제정한 BT.500 권고안의 주요 내용을 심층적으로 분석하여, 영상 화질에 대한 주관적 품질 평가 절차, 관찰자 선별 기준, 통계적 분석 방법, 및 신뢰구간 산출 기법 등을 체계적으로 고찰하고, 이를 토대로 UHDTV 환경에 적합한 화질 평가 방법론과 품질 측정 기준(안) 수립을 위한 기초 자료를 제공하는 것을 목적으로 한다.

2 주요 내용 요약

이 기술보고서는 5 장에서 ITU-R BT.500 권고안의 적용 범위 및 사용법에 대한 조언과 개정이력 등을 기술한다. 6 자에서는 주관적 화질 평가를 요구사항을 기술하고, 7 장에서는 ITU-R BT.500 에서 제시하고 있는 7 가지의 주관적 화질 평가 방법론을 설명한다. 8 장은 응용분야에 특화된 주관적 평각 방법론 중에서 SDTV, HDTV, 다중 프로그램 서비스에 대한 평가 방법론에 대해 살펴보고 9 장에서 결론을 기술한다.

3 인용 표준과의 비교

3.1 인용 표준과의 관련성

해당사항 없음.

3.2 인용 표준과 본 기술보고서의 비교표

해당사항 없음.

Preface

1 Purpose

This technical report presents a comprehensive analysis of the ITU-R BT.500 Recommendation established by the International Telecommunication Union - Radiocommunication Sector (ITU-R). It systematically examines the procedures for subjective video quality assessment, observer selection criteria, statistical analysis methods, and confidence interval estimation techniques. Based on these analyses, the report aims to provide foundational reference materials for developing quality evaluation methodologies and measurement standards applicable to UHDTV environments.

2 Summary

This technical report is organized as follows: Chapter 5 describes the scope of application of the ITU-R BT.500 Recommendation, provides guidance on its usage, and outlines its revision history. Chapter 6 presents the requirements for conducting subjective video quality assessments. Chapter 7 explains the seven subjective quality assessment methodologies proposed in ITU-R BT.500. Chapter 8 examines application-specific subjective assessment methods, focusing on evaluation methodologies for SDTV, HDTV, and multi-program services.

3 Relationship to Reference Standards

N/A

목 차

ITU-R BT.500 주관적 영상 품질 평가 기술 분석

(Analysis of Subjective Video Quality Assessment Technologies Based on ITU-R BT.500)

1 적용 범위

이 기술보고서는 국제 권고 ITU-R BT.500 의 주관적 화질 평가를 위한 요구사항 및 주관적 화질 평가 방법론을 검토·분석하고, UHDTV 품질 평가 기준안 수립을 위한 기초 자료 마련을 위해 SDTV, HDTV 등 응용분야에 특화된 주관적 평가 방법론을 살펴본다.

2 인용 표준

해당사항 없음.

3 용어 정의

해당사항 없음.

4 약어

BTC	Basic Test Cell
CBR	Constant Bit Rate
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double-Stimulus Impairment Scale
DVD	Design Viewing Distance
EBU	European Broadcasting Union
ESQS	Equivalent Single Quality Score

EVP	Expert Viewing Protocol
FPD	Flat Panel Display
HDR	High Dynamic Range
HDTV	High-Definition Television
JND	Just-Noticeable Differences
LDTV	Low-Definition Television
MOS	Mean Opinion Score
NTSC	National Television System Committee
OVD	Optimal Viewing Distance
PVD	Preferred Viewing Distance
PVS	Processed Video Sequences
QoE	Quality of Experience
SAMVIQ	Subjective Assessment Method for Video Quality in multimedia application
SC	Stimulus Comparison method
SDR	Standard Dynamic Range
SDSCE	Simultaneous Double Stimulus for Continuous Evaluation
SDTV	Standard-Definition Television
SIF	Standard Image Format
SOV	Segment Of Votes
SRC	Source Reference Sequences
SS	Single Stimulus method
SSCQE	Single Stimulus Continuous Quality Evaluation
SSMR	Single Stimulus with Multiple Repetition
SSNCS	SS procedure using an 11-grade Numerical Categorical Scale
UHDTV	Ultra High-Definition Television
VBR	Variable Bit Rate

5 ITU-R BT.500 권고안 개요

ITU-R BT.500 은 영상 시스템의 주관적 화질 평가(subjective assessment of video quality)를 수행하기 위한 국제 표준 권고로서, 다양한 영상 서비스 환경에서 화질 평가 결과의 일관성, 신뢰성 및 재현성을 확보하기 위해 필요한 절차, 조건 및 평가 기법을 규정한다. 본 권고는 수십 년간의 실험적 검증과 국제적 합의를 기반으로 발전해 왔으며, 영상 압축 코덱, 방송 전송 시스템, 디스플레이 장치 등 다양한 기술 분야에서 기준으로 활용되고 있다.

BT.500 의 핵심 목적은 다음과 같다.

- 주관적 화질 평가의 재현성과 신뢰성 확보
  서로 다른 기관, 장비, 환경에서도 동일한 절차를 통해 일관된 결과를 얻도록 표준화한다.
- 평가 조건의 통제 및 객관적 비교
  조명, 디스플레이 특성, 관찰 거리, 영상 소스 특성 등 실험 변수들을 명확히 규정한다.
- 다양한 평가 시나리오 적용성 확보
  방송, 스트리밍, 코덱 비교, 알고리즘 성능 검증 등 다양한 분야에서 활용될 수 있도록 여러 평가 방법론을 제안한다.
-

BT.500 은 크게 아래와 같이 3 개 파트로 구성된다.

- Part 1: 주관적 화질 평가를 위한 일반 요구사항
  - 실험 환경, 시청 조건, 평가 대상 신호, 관찰자 요건 등
- Part 2: 주요 평가 방법 및 절차
  - DSIS, SS 등 다양한 실험 기법
- Part 3: 응용 분야별 주관적 화질 평가 가이드라인
  - SDTV, HDTV, 멀티프로그램 환경 등 시나리오 평가 요건 및 지침

이러한 구조는 주관적 평가가 수행되는 전 과정(실험 준비 → 평가 절차 → 결과 분석)을 포괄적으로 포함하여, 실제 운용 환경에서의 적용 가능성을 높인다.

5.1 ITU-BT. 500 권고안 기반 본 보고서의 구성

이 보고서의 6, 7, 8 장은 BT.500 의 파트별 내용을 참고하여 아래와 같이 구성하였다.

6 장은 BT.500 의 Part 1 에 대응되는 섹션으로, 주관적 평가 수행 시 반드시 준수해야 하는 실험 환경 및 조건의 기본 규정을 설명한다. 이는 평가 결과의 신뢰성과 반복성을 확보하기 위한 기반으로, 다음 내용들을 포함한다.

- 시청 조건: 조명 환경, 시청 거리, 디스플레이 특성
- 소스 신호: 원본 콘텐츠 조건, 전송 파라미터
- 테스트 자료: 평가용 영상의 구성 방식
- 조건 범위 및 앵커링: 비교 기준 설정 방식
- 관찰자 요건: 시력 조건, 패널 수
- 테스트 세션 구성: 실험 운영 절차
- 결과 제시 형식: 데이터 분석 및 보고 형식

7 장은 BT.500 에서 규정하는 대표적인 평가 방법들을 설명한다. 화질 평가를 수행하는 실질적 실험 프로토콜을 제공하며, 실험 목적과 콘텐츠 특성에 따라 적합한 방법을 선택할 수 있도록 한다.

- DSIS: 부정적 영향 평가(Negative Impairment)
- DSCQS: 두 신호 간의 상대 비교(Double Stimulus)
- SS/SC: 단일 장면 기반의 단일 자극 평가(Single Stimulus)
- SSCQE: 스트리밍이나 시간 변화가 있는 콘텐츠에 적합한 연속 시간 평가 방식
- SAMVIQ, SDSCE, EVP 등 확장 기법

BT.500 은 다양한 방송·영상 서비스에 적용될 수 있도록 SDTV, HDTV, 멀티프로그램 환경 등 응용 시나리오별 지침을 추가적으로 제공한다. 8 장은 BT.500 의 이러한 응용 분야별 권고사항을 기반으로 다음을 설명한다.

- SDTV 환경에서의 주관적 평가 고려사항
- HDTV 환경에서의 평가 요건
- 멀티프로그램 환경(여러 채널 동시 인코딩/전송)에서의 평가 지침

6 주관적 화질 평가를 위한 요구사항

주관적 영상 평가 방법은 시청자의 반응을 보다 직접적으로 예측하는 측정법을 사용하여 텔레비전 시스템의 성능을 확인하는 데 사용된다. 이러한 점에서 시스템의 성능을 객관적인 수단만으로 완전히 규정하는 것이 불가능할 수 있으므로, 따라서 객관적인 측정을 주관적인 측정으로 보완할 필요가 있다.

일반적으로 주관적 평가는 두 가지 종류가 나눌 수 있다. 첫 번째는 최적의 조건 하에서 시스템의 성능을 확인하는 평가로, 일반적으로 품질 평가(Quality Assessments)라고 한다. 두 번째는 전송이나 송출과 관련된 비최적 조건 하에서 시스템이 품질을 유지하는 능력을 확인하는 평가로, 일반적으로 손상 평가(Impairment Assessments)'라고 한다.

가장 적절한 주관적 평가를 수행하기 위해서는, 먼저 사용 가능한 여러 선택지 중에서 요구되는 특정 상황과 영상 평가 목적에 가장 적합한 방법론을 선택해야 한다.

이러한 선택을 위해서는 이 장에 상세히 기술된 일반적인 특징을 고려하여 평가 대상의 문제나 프로세스에 가장 적합한 선택지가 무엇인지 파악해야 한다.

적합한 선택지들이 파악되면, 이 장의 후반부에서는 사용되는 평가자의 유형과 평가 환경의 상황을 고려하여 평가 대상의 문제나 프로세스에 가장 적합한 방법론을 선택하는 데 도움이 될 수 있는 권장된 영상 평가 방법론의 개요를 제공한다.

그럼에도 불구하고 가장 적절한 방법론의 선택은 평가 대상 시스템이 달성하고자 하는 서비스 목표에 따라 달라진다. 따라서 특정 응용 분야의 전체 평가 절차가 별도로 기술되어 있다.

이 장에서는 주관적 평가를 위한 일반적인 시청 조건을 제시한다. 특정 시스템의 주관적 평가를 위한 구체적인 시청 조건은 관련된 방법론에 기술되어 있다. 참고 – HDR 영상을 주관적으로 평가할 때에는, 해당 섹션에서 참조로 제시한 문서가 있는 경우 해당 문서를 참고하는 것이 권장된다[1].

[1] HDR 에 대한 추가적인 연구와 경험이 쌓임에 따라, 이 권고안은 추가적인 지침을 포함하도록 개정될 예정임.

6.1 일반 시청 조건

실험실 시청 환경은 시스템을 검증하기 위한 엄격한 조건을 제공하는 것을 목표로 한다. 실험실 환경에서의 주관적 평가를 위한 일반 시청 조건은 6.1.1 장에 기술되어 있다.

가정 시청 환경은 TV 체인의 소비자단에서 품질을 평가하는 수단을 제공하는 것을 목표로 한다. 6.1.2 장의 일반 시청 조건은 가정 환경을 재현한다. 이 매개변수들은 일반적인 가정 시청 상황보다 약간 더 엄격한 환경을 정의하도록 선정되어 있다.

6.1.1 실험실 환경에서의 주관적 평가를 위한 일반 시청 조건

평가자의 시청 조건은 다음과 같이 구성되어야 한다.

a)	실내 조도	낮음 (low)
b)	배경의 색도	D_{65}
c)	최대 휘도[2]	70-250 cd/m^2 (6.1.6.5 참조)
d)	디스플레이 명암비	≤ 0.02 (6.1.6.4 참조)
e)	영상 디스플레이 뒤 배경의 휘도 대 영상 최대 휘도 비율	≈ 0.15

6.1.2 가정 환경에서의 주관적 평가를 위한 일반 시청 조건

a)	스크린 위 환경 조도 (주변 환경에서 스크린에 비치는 입사광, 스크린에 수직으로 측정해야 함)	200 lux
b)	최대 휘도	70-500 cd/m^2 (6.1.6.4 참조)

[2] 최대 휘도는 실내 조도에 따라 조정해야 한다.

c) 비활성 스크린의 휘도 대 최대 휘도의 비율 (디스플레이 명암비) ≤ 0.02 (6.1.6.4 참조)

6.1.3 시청 거리

시청 거리는 스크린 크기에 기반하며, PVD 와 DVD 라는 두 가지 별개의 기준에 따라 선택될 수 있다. 두 기준 중 어느 하나를 선택할지는 연구의 목적에 따라 달라진다.

6.1.3.1 선호 시청 거리

선호 시청 거리는 경험적으로 결정된 시청자의 선호도에 기반한다. PVD (스크린 크기의 함수)는 그림 6-1 에 나와 있으며, 여기에는 참조 가능한 자료에서 수집된 다수의 데이터 세트가 포함되어 있다. 이 정보는 주관적 평가 테스트를 설계할 때 참조할 수 있다.

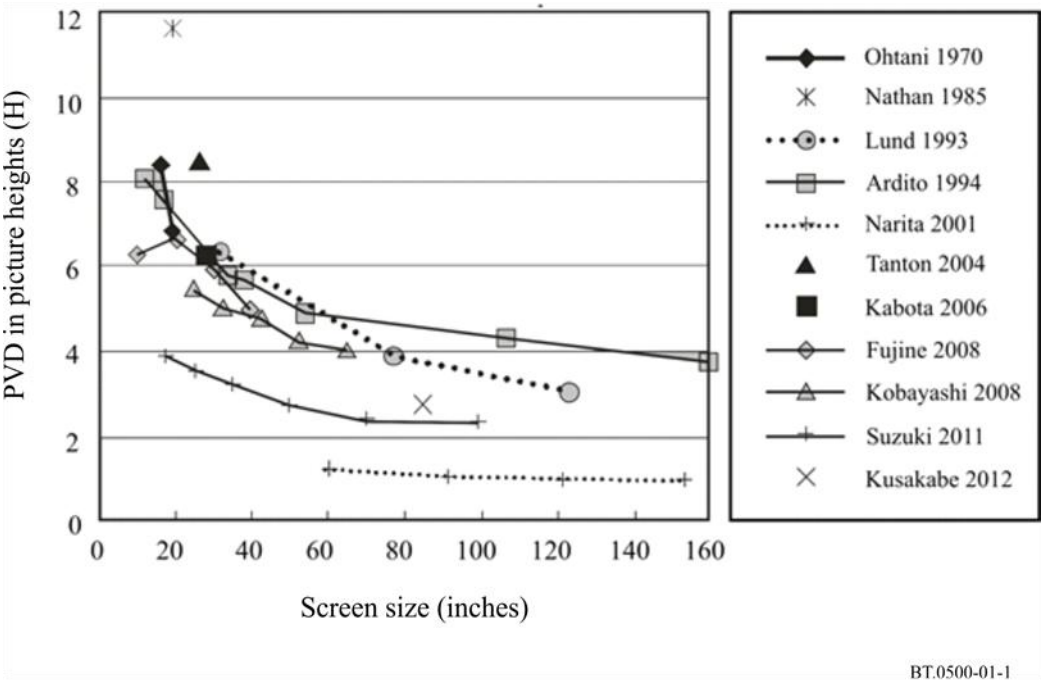


그림 6-1. 화면 크기에 따른 선호 시청 거리

Preferred viewing distance in function of the screen sizes

6.1.3.2 설계 시청 거리

디지털 시스템의 DVD 또는 OVD 란, 두 개의 인접한 픽셀이 관찰자의 눈에서 1 분각(arc-min)의 시각을 이루는 거리로 정의된다. 또한 최적 수평 시청 시야각(Optimal horizontal viewing angle)은 이러한 최적 시청 거리에서 영상을 관찰할 때 형성되는 각도를 의미한다.

표 6-1 은 여러 영상 해상도 시스템에 대한 최적 시청 거리(및 최적 수평 시야각)를 영상 높이의 배수로 표현하여 보여준다.

표 6-1. 최적 수평 시야각, 영상 높이(H)별 최적 시청 거리
Optimal horizontal viewing angle, optimal viewing distance in image heights(H)

Image system	Reference	Aspect ratio	Pixel aspect ratio	Optimal horizontal viewing angle	Optimal viewing distance
720 × 483	ITU-R BT.601	4:3	0.89	11°	7 *H*
640 × 480	VGA	4:3	1	11°	7 *H*
720 × 576	ITU-R BT.601	4:3	1.07	13°	6 *H*
1024 × 768	XGA	4:3	1	17°	4.5 *H*
1280 × 720	ITU-R BT.1543 and BT.1874	16:9	1	21°	4.8 *H*
1400 × 1050	SXGA+	4:3	1	23°	3.3 *H*
1920 × 1080	ITU-R BT.709	16:9	1	31°	3.2 *H*
3840 × 2160	ITU-R BT.2020	16:9	1	58°	1.6 *H*
7680 × 4320	ITU-R BT.2020	16:9	1	96°	0.8 *H*

참고: 영상 평가에 해상도가 포함되는 경우, 7680 × 4320 및 3840 × 2160 형식에 대해서는 더 낮은 값의 시청 거리를 사용해야 한다. 해상도가 평가 대상이 아닐 경우, (3840 × 2160 형식: 영상 높이의 1.6 ~ 3.2 배, 7680 × 4320 형식: 영상 높이의 0.8 ~ 3.2 배) 범위 내의 모든 시청 거리를 사용할 수 있다.

6.1.4 관찰 각도

화면에서 재현되는 색상의 변화가 관찰자에게 보이지 않도록, 법선(normal)에 대한 최대 관찰 각도는 제한되어야 한다. 또한 시험 중인 영상 시스템의 최적 수평 시청 각도를 고려하여 관찰 각도를 결정해야 한다. 자세한 내용은 보고서 ITU-R BT.2129 의 1.8 장을 참조하면 된다.

6.1.5 실내 환경 - 색상 구성

디스플레이 배경의 색상은 기준 백색점(reference white point)과 동일해야 하며, 나머지 실내 표면은 어두운 무광 재질을 사용해야 한다. 이는 디스플레이 화면에 도달하는 산란광(stray light)을 최소화하기 위함이다.

6.1.6 디스플레이

특성이 서로 다른 디스플레이를 사용할 경우 주관적 화질 평가 결과도 달라질 수 있다. 따라서 평가에 사용되는 디스플레이의 특성을 사전에 확인하는 것이 강력히 권장된다. 전문용 평판 디스플레이(FPD: Flat Panel Display)를 주관적 평가에 사용하는 경우, ITU-R BT.1886(HDTV 제작용 평판 디스플레이 기준 EOTF) 및 ITU-R BT.2129(HDTV 프로그램 제작 환경에서 마스터 디스플레이로서의 FPD 사용자 요구사항)를 참고할 수 있다.

보고서 ITU-R BT.2390 은 HDR 영상 평가를 위한 실험실 및 가정용 디스플레이와 시청 환경에 관한 정보를 제공한다.

6.1.6.1 디스플레이 처리

영상 스케일링, 프레임 레이트 변환, 이미지 향상과 같은 디스플레이 내부 처리 기능이 구현된 경우, 아티팩트(artefact)가 발생하지 않도록 처리해야 한다. HDR 처리 또한 평가 중인 HDR 시스템 또는 사용되는 HDR 형식에 적합해야 한다. 소비자 환경 또는 배포 평가에서는 정적 또는 동적 메타데이터 사용을 포함할 수 있다. 다른 실험실에서 평가를 정확하게 재현할 수 있도록 해당 메타데이터의 전체 세부 정보를 평가 기록에 포함해야 한다.

소비자용 디스플레이를 사용하여 주관적 영상 평가를 수행할 때는 해당 영상 처리의 영향이 평가의 대상이 아니라면, 모든 영상 처리 옵션을 비활성화하는 것이 중요하다.

인터레이스(interlace) 영상을 평가할 때에는 평가 보고서에 디인터레이서(de-interlacer) 사용 여부를 명시해야 한다. 인터레이스 신호를 디인터레이서 없이 표시할 수 있다면 사용하지 않는 것이 바람직하다.

6.1.6.2 디스플레이 해상도

전문가용 디스플레이는 일반적으로 지정된 휘도 영역 내에서 주관적 평가 기준에 부합하는 해상도를 제공한다.

사용된 휘도 값에서 디스플레이의 중앙 및 모서리 영역의 최대·최소 해상도를 확인하고 보고하는 것이 권장된다.

소비자용 FPD TV 를 사용하는 경우에도 동일하게 중앙 및 모서리의 최대·최소 해상도를 확인하고 보고하는 것이 강력히 권장된다.

현재 디스플레이 또는 소비자용 TV 의 해상도를 확인하기 위해 주관적 평가 수행자가 사용할 수 있는 가장 실용적인 방법은 전자적으로 생성된 스윕 테스트 패턴(swept test pattern)을 사용하는 것이다.

6.1.6.3 디스플레이 조정

디스플레이의 밝기와 명암비는 주변 조도에 맞추어 ITU-R BT.814 권고안의 PLUGE 파형을 사용해 조정해야 한다.

SDR 영상 평가 시에는 ITU-R BT.815 권고안에 따라 디스플레이의 명암비를 측정해야 한다. HDR 영상 평가 시에는 ITU-R BT.2390 보고서를 참조해야 한다.

6.1.6.4 디스플레이 명암비

디스플레이의 명암비는 주변 조도에 큰 영향을 받는다.

전문가용 디스플레이는 일반적으로 높은 조도 환경에서 명암비를 향상시키는 기술을 사용하지 않기 때문에, 높은 조도 환경에서는 요구되는 대비 기준을 충족하지 못할 가능성이 있다.

반면 소비자용 디스플레이는 높은 조도 환경에서 더 나은 명암비를 얻기 위한 기술을 일반적으로 사용한다.

6.1.6.5 디스플레이 밝기

LCD 디스플레이의 밝기를 조정할 때에는 신호 레벨 스케일링이 아니라 백라이트 강도 조절을 사용하는 것이 비트 정밀도를 유지하는 데 유리하다. 백라이트를 사용하지 않는 다른 디스플레이 기술의 경우, 신호 레벨 스케일링 이외의 방식으로 백색 레벨을 조정해야 한다. PDP 의 경우, 밝기는 광 방출 횟수로 제어되며, 밝기를 낮추면 톤 재현성이 저하될 수 있다는 점에 유의해야 한다.

6.1.6.6 디스플레이 모션 아티팩트

디스플레이 자체의 기술적 특성으로 인해 모션 아티팩트가 추가로 발생해서는 안된다. 반면, 입력 신호에 포함된 모션 효과는 디스플레이에 정확히 재현되어야 한다. 소비자용 디스플레이를 사용할 경우 모든 모션 처리 옵션을 반드시 비활성화해야 한다.

6.1.6.7 와이드스크린 16:9 화면비 디스플레이의 안전 영역

16:9 디스플레이의 안전 영역은 ITU-R BT.1848 권고안에 제시되어 있다.

6.2 소스 신호

소스 신호는 기준 이미지를 직접 제공하며, 테스트 대상 시스템의 입력이 된다. 이 신호는 사용되는 텔레비전 표준에 맞는 최적의 품질이어야 한다. 안정적인 결과를 얻기 위해서는 프레젠테이션 페어의 기준 부분에 결함이 없어야 한다.

디지털로 저장된 정지 이미지와 비디오 시퀀스는 재현성이 가장 높으므로 선호되는 소스 신호이다. 시스템 비교를 더 의미 있게 만들기 위해 실험실 간에 교환할 수 있다.

테스트 대상 시스템의 성능이 신호 이력의 초기 단계에서 수행되었을 수 있는 모든 처리 효과에 의해 어떻게 영향을 받는지 고려해야 하는 경우가 자주 있다. 따라서 비록 보이지 않더라도 처리 왜곡을 유발할 수 있는 체인의 일부 구간에 대해 테스트를 수행할 때마다, 결과 신호를 투명하게 기록하고, 연쇄적인 처리로 인한 손상이 체인을 따라 어떻게 누적되는지 확인하고자 할 때 후속 다운스트림 테스트에서 사용할 수 있도록 하는 것이 바람직하다. 이러한 녹화물은 필요시 향후 사용을 위해 테스트 자료 라이브러리에 보관해야 하며, 녹화된 신호의 이력에 대한 상세한 설명을 포함해야 한다. 필요한 경우 35mm 슬라이드 스캐너를 정지 이미지 소스로 사용할 수 있다. 사용 가능한 해상도는 기존 텔레비전 평가에 적합하다. 필름의 색상 측정(colorimetry) 및 기타 특성은 스튜디오 카메라 이미지와 다른 주관적인 외관을 제공할 수 있다. 이것이 결과에 영향을 미치는 경우, 종종 훨씬 덜 편리하지만 직접적인 스튜디오 소스를 사용해야 한다. 일반적으로 슬라이드 스캐너는 실제 상황이 그러하므로, 최상의 주관적 이미지 품질을 위해 이미지별로 조정해야 한다.

다운스트림 처리 용량 평가는 종종 컬러 매트(colour-matte)를 사용하여 이루어진다. 스튜디오 작업에서 컬러 매트는 스튜디오 조명에 매우 민감하다. 따라서 평가는 일관되게 고품질 결과를 제공하는 특수 컬러 매트 슬라이드 페어를 사용하는 것이 바람직하다. 필요한 경우 전경 슬라이드에 움직임을 도입할 수 있다.

6.3 테스트 자료

텔레비전 평가에 필요한 테스트 자료의 종류를 확립하는 데에는 여러 접근 방식이 있지만 실제로는 특정 평가 문제를 다루기 위해 특정 종류의 테스트 자료를 사용해야 한다. 일반적인 평가 문제와 이러한 문제를 다루기 위해 사용되는 테스트 자료의 개요가 표 6-2 에 나와 있다.

표 6-2. 테스트 자료의 선택[3]
Selection of test material

평가 문제	사용된 자료
평균적인 자료에 대한 전반적인 성능	일반적, “중요하지만 과도하지 않은”
용량, 중요 애플리케이션 (예: 기여, 후처리 등)	테스트된 애플리케이션에 대해 매우 중요한 자료를 포함하는 범위
“적응형” 시스템의 성능	사용된 “적응형” 방식에 매우 중요한 자료
약점 및 개선 가능성 식별	중요하고, 속성별로 특화된 자료
시스템이 가변적으로 보이는 요인 식별	매우 풍부하고 광범위한 자료
서로 다른 표준 간의 변환	차이점에 대해 중요 (예: 필드 레이트)

일부 파라미터는 대부분의 이미지나 시퀀스에 대해 유사한 수준의 손상을 유발할 수 있다. 이러한 경우, 매우 적은 수의 이미지나 시퀀스(예: 2 개)로 얻은 결과도 여전히 의미 있는 평가를 제공할 수 있다.

그러나 새로운 시스템은 장면이나 시퀀스 콘텐츠에 크게 의존하는 영향을 미치는 경우가 많다. 이러한 경우, 전체 프로그램 시간에 대해 손상 확률과 이미지 또는 시퀀스 콘텐츠의 통계적 분포가 존재하게 된다. 일반적으로 이 분포의 형태를 알 수 없으므로, 테스트 자료의 선택과 결과 해석은 매우 신중하게 이루어져야 한다.

일반적으로 결과를 해석할 때 고려할 수 있기 때문에 중요한(critical) 자료를 포함하는 것이 필수적이지만, 중요하지 않은 자료로부터 추정하는 것은 불가능하다. 장면이나 시퀀스 콘텐츠가 결과에 영향을 미치는 경우, 테스트 대상 시스템에 대해 “중요하지만 과도하지 않은(critical but not unduly so)” 자료를 선택해야 한다. “과도하지 않은”이라는 문구는 해당 이미지가 일반적인 프로그램 시간의 일부를 구성할 수 있음을 의미한다. 이러한 경우 최소 4 개의 항목을 사용해야 한다. 예를 들어, 절반은 명백히 중요한 것이고, 절반은 적당히 중요한 것이다.

[3] 모든 테스트 자료는 텔레비전 프로그램 콘텐츠의 일부가 될 수 있는 것이다. 테스트 자료 선택에 대한 추가 지침은 ITU-R BT.500 파트 1 의 부속서 3 과 4 를 참조하십시오.

6.3.1. ITU-R 테스트 시퀀스

여러 기관에서 테스트용 정지 이미지와 시퀀스를 개발해 왔다. 보고서 ITU-R BT.2245 는 이미지 품질 평가를 위한 HDR-TV 테스트 자료를 포함한 HDTV 및 UHDTV 에 대해 자세히 설명하고, 주관적 평가에 사용할 수 있는 HDTV 및 UHDTV 테스트 자료의 세부 정보를 제공한다. 테스트 자료 선택에 대한 추가 아이디어는 ITU-R BT.500 파트 1 의 부속서 1 과 2 에 나와 있다.

6.4 조건의 범위와 앵커링(anchoring)

대부분의 평가 방법은 관찰되는 조건의 범위와 분포 변화에 민감하므로, 평가 세션에는 변화되는 요인들의 전체 범위가 포함되어야 한다. 그러나 전체 범위를 모두 포함하는 것이 어려운 경우, 척도의 극단값에 해당하는 조건 일부를 함께 제시함으로써 보다 제한된 범위로 이를 대체할 수 있다. 이러한 조건들은 예시 형태로 제시되며 가장 극단적인 조건임을 명시적으로 알릴 수도 있고(직접 앵커링), 세션 전체에 분산 배치하되 극단값임을 명시하지 않을 수도 있다(간접 앵커링).

6.5 관찰자

관찰자는 평가의 목적에 따라 전문가 또는 비전문가일 수 있다. 전문가 관찰자는 테스트 대상 시스템에 의해 발생할 수 있는 이미지 아티팩트에 대한 전문 지식을 가진 관찰자이다. 비전문가(“순수”) 관찰자는 테스트 대상 시스템에 의해 발생할 수 있는 이미지 아티팩트에 대한 전문 지식이 없는 관찰자이다. 어떤 경우에도 관찰자는 테스트 대상 시스템의 개발에 직접 관여했거나, 즉 특정하고 상세한 지식을 습득할 만큼 충분히 관여해서는 안 된다.

6.5.1. 관찰자의 수

선택한 방법론에서 달리 명시하지 않는 한, 최소 15 명의 관찰자를 사용해야 한다. 필요한 평가자의 수는 채택된 테스트 절차의 민감도와 신뢰도, 그리고 예상되는 효과의 크기에 따라 달라진다. 탐색적 성격과 같이 범위가 제한된 연구의 경우 15 명 미만의

관찰자를 사용할 수 있다. 이 경우, 해당 연구는 '비공식적'인 것으로 명시해야 한다. 관찰자들의 텔레비전 이미지 품질 평가에 대한 전문성 수준을 보고해야 한다.

6.5.2 관찰자 선별

일반적으로 세션에 앞서, 관찰자들은 Snellen 또는 Landolt 시력표를 통해 (교정) 정상 시력을, 그리고 특별히 선택된 차트(예: 이시하라)를 사용하여 정상 색각을 확인하기 위해 선별되어야 한다.

6.7.2.3 과 6.7.2.4 는 다양한 테스트 방법론에 적용될 수 있는 여러 관찰자 선별 시나리오를 상세히 설명한다. 여러 장소 또는 여러 기관의 테스트 프로그램을 일부로 하여 덜 공식적인 테스트를 수행하는 실험실의 경우, 관찰자 선별 방법과 기준의 전체 세부 정보를 교환하고 발표된 결과의 일부로 포함하는 것이 중요하다.

일반적으로 평가 패널의 특성(예: 방송 기관 직원, 대학생, 사무직 근로자 등 직업 범주, 성별, 연령대)에 대해 가능한 한 많은 세부 정보를 포함해야 한다.

참고 – 여러 테스트 실험실 간의 일관성에 대한 한 연구에서 서로 다른 실험실에서 얻은 결과 간에 체계적인 차이가 발생할 수 있음을 발견했다. 이러한 차이는 실험의 신뢰성과 재현성을 향상시키기 위해 여러 다른 실험실의 결과를 취합하려 할 때 특히 중요하다.

여러 실험실 간의 차이에 대한 한 가지 가능한 설명은 서로 다른 관찰자 그룹 간에 다른 기술 수준이 존재할 수 있다. 이 가설의 타당성을 평가하고, 입증될 경우 이 요인에 의해 기여된 변동을 정량화하기 위해 추가 연구가 수행되어야 한다.

6.5.3 평가 지침

평가자에게 평가 방법, 발생 가능성이 있는 손상 또는 품질 요인의 유형, 채점 척도, 순서 및 타이밍을 신중하게 소개해야 한다. 평가할 손상의 범위와 유형을 시연하는 훈련 시퀀스를 사용해야 한다. 이때 사용되는 예시 이미지는 테스트에 사용되는 이미지와는 다르지만 민감도는 비슷한 것이어야 한다. 품질 평가의 경우, 품질은 특정 지각적 속성으로 구성된 것으로 정의될 수 있다.

6.6 테스트 세션

한 세션은 30 분 이상 지속되어서는 안 된다. 첫 세션 시작 시, 관찰자의 의견을 안정시키기 위해 약 5 개의 “더미 프레젠테이션”을 도입해야 한다. 이 프레젠테이션에서 나온 데이터는 테스트 결과에 포함되어서는 안 된다. 여러 세션이 필요한 경우, 다음 세션의 시작 시에는 약 3 개의 더미 프레젠테이션만 필요하다.

프레젠테이션에는 무작위 순서(예: 그레코-라틴 방진에서 파생)를 사용해야 하지만, 피로나 적응이 채점에 미치는 영향을 세션 간에 상쇄할 수 있도록 테스트 조건 순서를 배열해야 한다. 일관성을 확인하기 위해 일부 프레젠테이션은 세션 간에 반복될 수 있다.

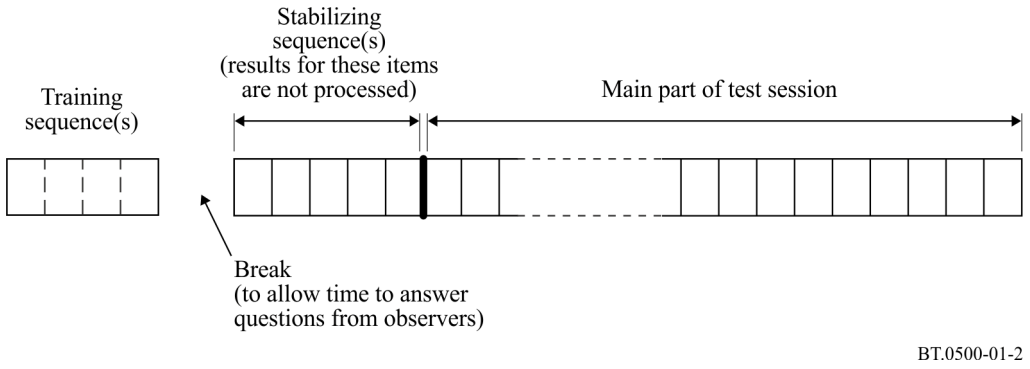


그림 6-2. 테스트 세션 구성

PRESENTATION STRUCTURE OF TEST SESSION

6.7 결과 제시

주관적 화질 평가 결과를 분석하여 제시하는 방법에 있어서 ITU-BT.500 권고안 파트 1 의 부속서 1 에 기술된 내용을 6.7.1~6.7.4 에 정리한다.

6.7.1 결과 분석 개요

텔레비전 시스템 성능을 평가하기 위한 주관적 실험 과정에서는 많은 양의 데이터가 수집된다. 이러한 데이터는 관찰자의 평가 점수표 또는 전자 형태로 존재하며, 이를 통계적 기법을 사용하여 시험 중인 시스템의 성능을 요약하는 그래프, 수치, 공식 또는 알고리즘 형태의 결과로 도출해야 한다.

본 절에서 기술되는 화질평가 결과분석 방법은 ITRU-BT.500 에서 제시되는 SS 방식, DSIS 방식, DSCQS 방식을 통한 TV 화질 평가 결과 뿐만 아니라 숫자 척도를 사용하는 기타 화질평가 방식들에 적용될 수 있다. SS 방식 및 DSIS 방식은 화질열화를 5 등급 또는 다중 등급 척도로 평가되며, DSCQS 방식은 연속적인 평가 척도로 이루어지며, 그 결과(참조 영상과 시험영상 간의 평가 점수 차이)는 0~100 사이의 정수값으로 정규화된다.

6.7.2 일반적인 분석 방법

각각의 화질평가 방식의 원칙에 따라 수행된 테스트는 1 과 5 사이 또는 0 과 100 사이와 같은 정수 값의 분포를 생성한다. 이러한 분포에는 관찰자 간의 판단 차이나, 여러 이미지 또는 시퀀스의 사용 등 실험에 관련된 다양한 조건의 영향으로 인해 변동이 발생할 수 있다.

한의 시험은 여러 번의 프레젠테이션(L)으로 구성되며, 각 프레젠테이션은 여러 개의 시험 시퀀스 또는 시험 이미지(K) 중 하나에 적용된 다수의 시험조건(J) 중 하나가 된다. 경우에 따라서 시험 시퀀스 또는 시험 이미지와 시험조건의 조합은 여러 번(R) 반복될 수 있다.

6.7.2.1 평균점수 계산

결과 분석의 첫번째 단계는 각 프레젠테이션에 대한 평균점수를 식(6-1)과 같이 계산하는 것이다.

$$\bar{u}_{jkr} = \frac{1}{N}\sum_{i=1}^{N} u_{ijkr} \qquad (6\text{-}1)$$

여기서,

- u_{ijkr}: 시험조건 j, 시험 시퀀스/영상 k, 반복 r 에 대해 관찰자 i가 부여한 점수
- N: 관찰자 수

마찬가지로, 각 시험조건 및 각 시험영상에 대해서도 전체 평균점수를 계산할 수도 있다.

6.7.2.2 신뢰구간 계산

6.7.2.2.1 원본(보상되지 않은 및/또는 근사되지 않은) 데이터 처리

시험 결과를 제시할 때 모든 평균점수에 대해 해당 샘플의 표준편차와 샘플 크기로부터 계산된 신뢰구간을 함께 제시해야 한다. 이 때, 95% 신뢰구간을 사용하는 것이 권장되며, 계산은 식(6-2)와 같다.

$$\left[\bar{u}_{jkr} - \delta_{jkr},\ \bar{u}_{jkr} + \delta_{jkr}\right] \qquad (6\text{-}2)$$

여기서,

$$\delta_{jkr} = 1.96\,\frac{S_{jkr}}{\sqrt{N}} \qquad (6\text{-}3)$$

각 프레젠테이션에 대한 표준편차는 식(6-4)와 같다.

$$S_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{(\bar{u}_{jkr} - u_{ijkr})^2}{(N-1)}} \qquad (6\text{-}4)$$

개별 점수들의 분포가 특정 요구 사항을 만족하는 조건 하에서, 95%의 확률로 실험적 평균점수와 '참' 평균점수(평가자 수가 매우 많은 경우) 간의 차이의 절댓값은 95% 신뢰구간보다 작다. 유사하게, 각 시험조건에 대해 표준편차를 계산할 수 있다. 그러나 시험영상의 수가 적은 경우에는 이 표준편차는 평가자 간의 차이보다 사용된 영상 간의 차이에 더 큰 영향을 받게 된다는 점에 유의해야 한다.

6.7.2.2.2 보정 및/또는 근사된 데이터의 처리

평가 척도 상의 잔여 손상/개선 및 경계 효과가 보정된 데이터나, 근사화 이후 왜곡 반응 또는 왜곡 가산 법칙 형태로 표현된 데이터의 경우(이는 실험에서 얻어진 품질 평균점수가 이러한 왜곡에 의존하기 때문), 신뢰구간은 해당 변수의 분산을 고려한

통계적 변수 변환을 사용하여 계산해야 한다. 또한 품질 평가 결과가 왜곡 반응 형태(즉, 실험 곡선)으로 제시되는 경우, 신뢰구간의 상한값과 하한값은 각 실험값에 대한 함수로 표현된다. 이러한 신뢰구간 한계를 계산하기 위해서는, 각 실험값에 대해 표준편차를 구하고, 그 표준편차가 실험값에 따라 어떻게 달라지는지를 근사적으로 평가해야 한다.

6.7.2.3 관찰자의 사후 선별

6.7.2.3.1 SSCQE 방식을 제외한 DSIS, DSCQS 및 대안 방식을 위한 첨도(Kurtosis) 기반 사후 선별

먼저, 각 프레젠테이션에 대한 점수 분포가 정규분포를 따르는지 여부를 β_2 검정(beta-2 test)을 이용해 확인해야 한다. 이 검정은 첨도계수(kurtosis coefficient), 즉 함수의 4 차 모멘트를 2 차 모멘트의 제곱으로 나눈 비율을 계산하는 방식이다. 만약 β_2 값이 2 이상 4 이하라면, 그 분포는 정규분포로 간주할 수 있다. 각 프레젠테이션에 대해 각 관찰자의 점수는 평균값에 표준편차의 배수(정규분포의 경우 2 배, 비정규분포의 경우 $\sqrt{20}$ 배)를 더한 값인 P_{jkr} 과 평균값에서 동일한 표준편차의 배수를 뺀 값인 Q_{jkr} 과 비교하여 관찰자의 점수가 P_{jkr} 보다 크면 해당 관찰자의 카운터 Pi 가 증가시키고, 관찰자의 점수가 Q_{jkr} 보다 작으면 Q_i 가 증가시킨다. 이 과정을 모든 프레젠테이션에 대해 수행한 뒤, 아래 두 비율을 계산하여 (1) 비율이 5%보다 크고, (2) 비율이 30%보다 작으면, 관찰자 i 는 실험에서 제외해야 한다.

(1) $(P_i + Q_i) / N_i$: 각 관찰자가 전체 평가 중 평균에서 과도하게 벗어난 횟수의 비율

(2) $|(P_i - Q_i) / (P_i + Q_i)|$: 평균보다 높게 또는 낮게 점수를 주는 편향 정도의 비율

참고 – 이 절차는 주어진 실험 결과에 두 번 이상 적용되어서는 안된다. 또한, 이 절차의 사용은 상대적으로 적은 수(예: 20 명 미만)의 관찰자가 있고 모두 비전문가인 경우로 제한되어야 한다. 이 절차는 EBU 방식(DSIS)에 권장되며, DSCQS 방식 및 다른 대안적 방식들에도 성공적으로 적용되었다.

위 과정은 수학적으로 식(6-5)와 같이 표현될 수 있다:

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2} \text{ with } \qquad m_x = \frac{\sum_{i=1}^{N} (u_{ijkr} - \bar{u}_{ijkr})^x}{N} \qquad (6\text{-}5)$$

$2 \le \beta_{2jkr} \le 4$ 이면,

$u_{ijkr} \ge \bar{u}_{jkr} + 2\,S_{jkr}$이면 Pi = Pi + 1

$u_{ijkr} \le \bar{u}_{jkr} - 2\,S_{jkr}$이면 Qi = Qi + 1

아니면,

$u_{ijkr} \ge \bar{u}_{jkr} + \sqrt{20}\,S_{jkr}$ 이면 Pi = Pi + 1

$u_{ijkr} \le \bar{u}_{jkr} - \sqrt{20}\,S_{jkr}$이면 Qi = Qi + 1

$\frac{P_i + Q_i}{J \cdot K \cdot R} > 0.05$ 이고 $\left|\frac{P_i - Q_i}{P_i + Q_i}\right| < 0.3$ 이면 관찰자 i 는 제외

여기서,

N: 관찰자 수

J: 시험조건 수

K: 시험영상 수

R: 반복 수

L: 프레젠테이션 수

6.7.2.3.2 SSCQE 방식을 위한 첨도(Kurtosis) 기반 사후 선별

SSCQE 방식의 시험 절차를 사용할 때의 특정 관찰자 선별에서는 적용 단위는 시험 구성(시험조건과 시험 시퀀스의 조합)이 아니라 시험 구성내의 일정한 시간구간(예: 10 초 투표 구간)이 된다. 이 절차는 2 단계의 필터링으로 이루어진다. 첫 번째 단계에서는 평균적인 평가 경향에 비해 점수가 현저히 치우친 관찰자를 탐지하고 제외하며, 두 번째 단계에서는 점수의 전반적인 치우짐과는 상관없이 평가가 일관되지 않거나 불규칙한 관찰자를 탐지하고 제외한다.

- 1 단계: 국소적 평가 역전 탐지

여기서도, 각 시험 구성의 각 시간구간에 대한 점수 분포가 정규분포를 따르는 지 여부를 β_2 검정을 이용하여 먼저 확인해야 한다. 만약 β_2 가 2 와 4 사이이면 해당 분포는 정규분포로 간주할 수 있다. 그런 다음, 이 절차는 각 시험 구성의 각 시간구간에 대해 식(6-6)에 따라 적용된다.

$$\beta_{2jklr} = \frac{m_4}{(m_2)^2} \text{ with } \qquad m_x = \frac{\sum_{n=1}^{N}(u_{njklr} - \bar{u})^x}{N} \qquad (6\text{-}6)$$

$2 \le \beta_{2jkr} \le 4$ 이면,

$u_{ijkr} \ge \bar{u}_{jkr} + 2\,S_{jkr}$ 이면 Pi = Pi + 1

$u_{ijkr} \le \bar{u}_{jkr} - 2\,S_{jkr}$ 이면 Qi = Qi + 1

아니면,

$u_{ijkr} \ge \bar{u}_{jkr} + \sqrt{20}\,S_{jkr}$ 이면 Pi = Pi + 1

$u_{ijkr} \le \bar{u}_{jkr} - \sqrt{20}\,S_{jkr}$ 이면 Qi = Qi + 1

$\frac{P_i}{J \cdot K \cdot L \cdot R} > X$ 또는 $\frac{Q_i}{J \cdot K \cdot L \cdot R} > X$ 이면 관찰자 i 는 제외

여기서,

N: 관찰자 수

J: 시험조건과 시험영상의 조합내에 포함되 시간구간 수

K: 시험조건 수

L: 시험영상 수

R: 반복 수

1 단계 절차는 평균점수에서 현저하게 벗어난 평가를 한 관찰자를 제외한다. 그림 6-1 은 이러한 예시 2 가지를 나타낸 것으로, 2 개의 극단적인 곡선은 평균으로부터 큰 편차를 보이는 경우이다. 하지만, 이 제외 기준만으로는 평가의 또 다른 중요한 편향의 원인인 역전 현상을 탐지하지 모하므로 두 번째 절차 단계가 제안된다.

- 2 단계: 국소적 평가 역전 탐지

2 단계에서는 탐지는 관찰자 선별 수식을 기반으로 수행된다. 단 적용 범위에 대해 약간의 수정이 이루어 진다. 입력 데이터 세트는 다시 모든 시험 구성의 모든 시간구간(예: 10 초)에 대한 점수로 구성된다. 그러나 이번에는 1 단계에서 이미 처리된 편차 효과를 최소화하기 위해 점수들을 전체 평균을 중심으로 사전에 주심화(centering)한다. 이후는 이전과 동일한 절차가 적용된다.

각 시험 구성의 각 시간구간에 대한 점수 분포가 정규분포를 따르는 지 여부를 β_2 검정을 이용하여 먼저 확인해야 한다. 만약 β_2 가 2 와 4 사이이면 해당 분포는 정규분포로 간주할 수 있다. 그런 다음, 이 절차는 각 시험 구성의 각 시간구간에 대해 식(6-7)에 따라 적용된다.

$$\bar{u}_{klr} = \frac{1}{N} \cdot \frac{1}{J} \sum_{n=1}^{N} \sum_{j=1}^{J} u_{njklr} \qquad (6\text{-}7)$$

마찬가지로, 각 시험 구성 및 각 관찰자에 대한 평균점수는 식(6-8)과 같이 정의된다.

$$\bar{u}_{nklr} = \frac{1}{J} \sum_{j=1}^{J} u_{njklr} \qquad (6\text{-}8)$$

여기서 $\bar{u}_{nklr}$ 은 관찰자 i 가 시간구간 j, 시험조건 k, 시험영상 l, 반복 r 에서 부여한 점수를 나타낸다.

각 관찰자에 대해 중심화된 점수 u^*_{njklr} 는 식(6-9)와 같이 계산된다.

$$u^*_{njklr} = u_{njklr} - \bar{u}_{nklr} + \bar{u}_{klr} \qquad (6\text{-}9)$$

각 시험 구성의 각 시간구간마다 평균 u^*_{njklr}, 표준편차 S^*_{jklr}, 첨도계수 $\beta_2{}^*_{jklr}$ 를 계산한다. 첨도계수 $\beta_2{}^*_{jklr}$ 는 식(6-10)으로 정의된다.

$$\beta_2{}^*_{jklr} = \frac{m_4}{(m_2)^2} \text{ with } \qquad m_x = \frac{\sum_{n=1}^{N} (u^*_{njklr})^x}{N} \qquad (6\text{-}10)$$

$2 \le \beta_2{}^*_{jklr} \le 4$ 이면,

$u^*_{njklr} \ge \bar{u}^*_{jklr} + 2\,S^*_{jklr}$ 이면 P*i = P*i + 1

$u^*_{njklr} \le \bar{u}^*_{jklr} - 2\,S^*_{jklr}$ 이면 Q*i = Q*i + 1

아니면,

$u^*_{njklr} \geq \bar{u}^*_{jklr} + \sqrt{20}\ S^*_{jklr}$ 이면 P*i = P*i + 1

$u^*_{njklr} \leq \bar{u}^*_{jklr} - \sqrt{20}\ S^*_{jklr}$ 이면 Q*i = Qi* + 1

$\frac{P^*_i + Q^*_i}{J \cdot K \cdot L \cdot R} > Y$ 이고 $\left|\frac{P^*_i - Q^*_i}{P^*_i + Q^*_i}\right| < Z$ 이면 관찰자 i는 제외

연기서,

N: 관찰자 수

J: 시험조건과 시험영상의 조합내에 포함된 시간구간 수

K: 시험조건 수

L: 시험영상 수

R: 반복 수

이 방법에 적합한 것으로 경험적으로 확인된 매개변수 X, Y, Z 의 제안값은 각각 0.2, 0.1, 0.3 이다.

6.7.2.3.3 상관관계 기반 사후 선별

각 관찰자는 각 장면과 알고리즘(시험조건)에 대해 품질 저하 정도를 공정하고 일관되게 평가해야 한다. 제외 기준은 특정 관찰자의 점수가 해당 시험 세션에서 모든 관찰자의 평균점수와 비교하여 얼마나 일관성 있게 유지되는지를 확인한다. 판단 기준은 각 관찰자의 개별 점수와 모든 관찰자의 평균점수 간의 상관관계를 기반으로 한다. 이 절차는 이전에 설명한 복잡한 방법에 비해 구현이 간단하다.

6.7.2.3.3.1 피어슨 상관관계

피어슨 상관관계를 적용하기 위해서는 품질 척도와 관찰자 점수 범위 간의 관계가 선형이라고 가정해야 한다. 이 절차의 주요 목적은 간단한 방법으로 특정 관찰자의 점수가 전체 관찰자들의 평균점수와 얼마나 일관성있게 일치하는지를 식 11 을 적용하여 검증하는 것이다. 숨겨진 기준 영상은 고품질 앵커(anchor)로 간주된다. 만약 평가에 저품질 앵커와 고품질 앵커가 모두 포함되어 있다면, 이는 전체 상관계수의 값을 높이는 경향이 있으며, 반대로 관찰자 간의 상관계수 차이는 감소하게 된다.

$$r(x,y)=\frac{\sum_{i=1}^{n}x_i y_i-\frac{\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}\right)}} \quad (6\text{-}11)$$

여기서,

- x_i: 특정 조합 *(알고리즘, 비트율, 장면)* 에 대한 모든 관찰자의 전체 평균점수
- y_i: 한 명의 특정 관찰자가 동일한 *(알고리즘, 비트율, 장면)* 조합에 대해 부여한 개별 점수
- n : 전체 데이터의 수 즉, (알고리즘의 개수) × (장면의 개수) x (비트율 수)
- i : 평가 인덱스로 각 데이터 포인트를 고유하게 식별하는(코덱번호, 비트율번호, 장면번호)의 조합

6.7.2.3.3.2 스피어만 순위 상관관계

식(6-12)와 같이 정의되는 스피어만 순위 상관계수는 품질 척도와 관찰자 점수 범위 간의 관계가 선형이 아니라고 가정되는 경우에도 적용할 수 있다. 일반적으로 피어슨 상관계수 결과와 스피어만 순위 상관계수 결과는 매우 유사하다.

$$r(x,y)=\left[1-\frac{6\times\sum_{i=1}^{n}\left[R(x_i)-R(y_i)\right]^2}{n^3-n}\right] \quad (6\text{-}12)$$

여기서,

- x_i: 특정 조합 *(알고리즘, 비트율, 장면)* 에 대한 모든 관찰자의 전체 평균점수
- y_i:한 명의 특정 관찰자가, 동일한 *(알고리즘, 비트율, 장면)* 조합에 대해 부여한 개별 점수

- n :전체 데이터의 수 즉, (알고리즘의 개수) × (장면의 개수) x (비트율 수)
- i : 평가 인덱스로 각 데이터 포인트를 고유하게 식별하는 (코덱 번호, 비트율 번호, 장면 번호)의 조합
- $R(x_i \text{ or } y_i)$: x_i 또는 y_i 대한 순위

6.7.2.3.3.3. 시험 관찰자 제외를 위한 최종 제외 기준

스피어만 순위 상관계수와 피어슨 상관계수는 다음 조건에 따라 관찰자를 제외하기 위해 계산된다.

IF [mean(r) – sdt(r)] > MCT

Rejection threshold = MCT

ELSE

Rejection threshold = [mean(r) – sdt(r)].

IF [r (Observer i)] > Rejection threshold.

관찰자 i 는 포함

ELSE

관찰자 i 는 제외

여기서,

- r = min (피어슨 상관계수, 스피어만 순위 상관계수)
- mean(r): 모든 관찰자의 상관계수 평균
- sdt(r): 상관게수의 표준편차
- Max Correlation Threshold (MCT) = 0.85

MCT 값 0.85 는 SAMVIQ 및 DSCQS 방식에 유효하며, 그 외의 경우(즉, SS 및 DSIS 방식)에는 MCT 값 0.7 을 적용해야 한다.

6.7.2.4 도전적인 시험조건 하에서의 평균점수 및 신뢰구간 계산

주관적 화질 평가는 주로 도전적인 조건 하에서 수행되어야 한다. 예를 들어, 크라우드소싱 시험에서는 참가자들이 실험실보다 덜 통제된 환경에서 평가를 수행하게 된다. 또한, 여러 실험실이 참여하는 대규모 시험의 경우에는 수집된 점수의 분산이 커질 수 있다. 이러한 상황에서는 6.7.2.1 에서 6.7.2.3 에 소개된 방법들은 적용되기 어렵다.

이 절에서는 복원된 평균점수와 신뢰구간의 데이터 품질을 개선하는 것으로 나타난 고급 데이터 분석 기법을 소개한다.

이 기법의 기본 개념은 다음과 같다. 각 평가자의 행동 특성을 명시적으로 모델링하는 것이 유용하다. 특히, 편향(bias)과 일관성(consistency)은 평가자의 점수에 영향을 미치는 대표적인 인간요인이다. 이 기법은 반복적인 절차를 통해, 각 프레젠테이션의 진짜 품질과 각 평가자의 편향 및 일관성을 동시에 추정한다. 이때 추정된 각 프레젠테이션의 진짜 품질은 편향이 제거되고 일관성으로 가중된 평균의견점수(MOS)로 해석될 수 있다. 이 방법은 6.7.2.3.1 에서 설명된 관찰자 사후 필터링 방식은 평가자를 전부 유지하거나 제외하는 강성 제외(hard rejection)하는 반면 이 방법은 일관성이 낮은 이상치 평가자의 경우 그 평가자의 점수는 낮은 가중치를 적용하여 전체 MOS 에 거의 영향을 끼치지 않는 연성 제외 (soft rejection)이라 할 수 있다.

이 기법의 부산물은 각 평가자의 편향과 일관성을 추정하는 것이다. 이는 평가자가 주관적 시험 수행에 적합한지에 대한 귀중한 정보이므로, 향후 평가 참여자 선정시 선별 기준으로 활용될 수 있다. 예를 들어, 평가자가 매우 일관성 없게 투표하는 것으로 나타나면, 향후 시험에서는 참여 제외 대상이 될 수 있다.

이 기법은 먼저 모든 평가자와 모든 반복에 걸쳐 각 프레젠테이션의 평균점수를 식(6-13)에 따라 추정한다.

$$\bar{u}_{jk} = \frac{1}{N \cdot R} \sum_{i=1}^{N} \sum_{r=1}^{R} u_{ijkr} \qquad (6\text{-}13)$$

여기서, u_{ijkr} 은 관찰자 i 가 시험조건 j, 시험영상 k, 반복 r 에 대해 부여한 점수를 의미한다. 또한, N 은 모든 관찰자 수, R 은 전체 반복 횟수를 나타낸다.

다음 단계에서, 각 평가자의 편향 b_i 는 식(6-14)에 의해 추정된다.

$$b_i = \frac{1}{J \cdot K \cdot R} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{r=1}^{R} u_{ijkr} - \bar{u}_{jk} \qquad (6\text{-}14)$$

여기서, J 는 시험조건의 개수, K 는 시험 시퀀스의 개수를 나타낸다. 이후에는 다음 단계들이 반복 루프형태로 수행된다.

각 프레젠테이션에 대한 현재 추정된 평균점수를 식(6-15)에 나타낸 $\bar{u}_{jk}^{c}$로 나타낸다.

$$\bar{u}_{jk}^{c} = \bar{u}_{jk} \qquad (6\text{-}15)$$

이어서, 각 관찰된 평가 점수에서 현재 추정된 평균점수와 관찰자 편향으로 설명되지 않는 잔차를 식(6-16)에 의해 계산한다.

$$e_{ijkr} = u_{ijkr} - \bar{u}_{jk} - b_i \qquad (6\text{-}16)$$

이러한 잔차들은 식(6-17)에 나타내 바와 같이 각 평가자의 비일관성 σ_i 를 계산하는데 사용된다.

$$\sigma_i = \sqrt{\frac{1}{J \cdot K \cdot R} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{r=1}^{R} (u_{ijkr} - \mu_{e_i})^2} \qquad (6\text{-}17)$$

여기서,

$$\mu_{e_i} = \frac{1}{J \cdot K \cdot R} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{r=1}^{R} e_{ijkr} \qquad (6\text{-}18)$$

다음으로 새로운 평균점수가 식(6-19)에 의해 계산될 수 있다.

$$\bar{u}_{jk} = \frac{\sum_{i=1}^{N} \sum_{r=1}^{R} \sigma_i^{-2} (u_{ijkr} - b_i)}{\sum_{i=1}^{N} \sum_{r=1}^{R} \sigma_i^{-2}} \qquad (6\text{-}19)$$

이어서, 식(6-12)에 따라 각 평가자의 편향을 갱신한다.

반복 루프는 식(6-20)을 만족하면 종료된다.

$$\sum_{j=1}^{J} \sum_{k=1}^{K} (\bar{u}_{jk} - \bar{u}_{jk}^{c})^2 \qquad (6\text{-}20)$$

반복 루프가 종료된 후, 각 프레젠테이션에 대한 점수의 표준편차는 식(6-21)에 의해 계산된다.

$$S_{jk} = \frac{\sigma_j}{\sqrt{N}} \qquad (6\text{-}21)$$

여기서,

$$\sigma_j = \sqrt{\frac{1}{N \cdot R} \sum_{i=1}^{N} \sum_{r=1}^{R} (e_{ijkr} - \mu_{e_j})^2} \qquad (6\text{-}22)$$

이고

$$\mu_{e_j} = \frac{1}{N \cdot R} \sum_{i=1}^{N} \sum_{r=1}^{R} e_{ijkr} \qquad (6\text{-}23)$$

이다.

최종 신뢰구간은 식(6-2)와 식(6-3)에 따라 계산된다.

6.7.3 영상 왜곡의 평균점수와 객관적 측정치 간의 관계를 찾기 위한 처리

만약 왜곡의 객관적인 측정과 식(6-1)에 의해 계산된 평균점수 사이의 관계를 조사하기 위해 주관적인 시험이 수행되었다면, 평균점수와 왜곡 파라미터 사이의 간단한 연속 관계를 찾는 것으로 구성된 다음 절차를 사용할 수 있다.

6.7.3.1 대칭 로지스틱 함수에 의한 근사

이 실험적 관계를 로지스틱 함수로 근사하는 것은 흥미롭다. 주관적 평균점수 $\bar{u}$의 데이터를 처리하는 절차는 다음과 같이 수행할 수 있다.

$\bar{u}$ 값의 척도는 연속변수 p 를 취하여 식(6-24)와 같이 정규화된다.

$$p = (\bar{u} - u_{min})/(u_{max} - u_{min}) \qquad (6\text{-}24)$$

여기서,

u_{min}: 최저 품질에 대한 u-척도에서 사용 가능한 최소 점수
u_{max}: 최고 품질에 대한 u-척도에서 사용 가능한 최대 점수

P 와 왜곡척도 D 사이의 관계를 그래프로 나타내면, 그 곡석은 비대칭적인 S 자형(sigmoid) 형태를 띠는 경향이 있음을 알 수 있다. 단 이는 D 값의 자연적인 범위가 $\bar{u}$ 가 급격히 변하는 구간(즉, 중간 품질 구간)을 충분히 벗어날 만큼 넓게 확장되어 있을 때 성립한다.

함수 $p = f(D)$ 는 적절하게 선택된 로지스틱 함수로 근사할 수 있으며, 그 일반적인 형태는 식(6-25)와 같다.

$$p = 1/[1 + \exp(D - D_M) \cdot G] \qquad (6\text{-}25)$$

여기서, D_M과 G 는 상수이며, G 는 양수 또는 음수일 수 있다.

최적의 로지스틱 함수 근사로부터 얻은 p 값을 사용하여 식(6-26)에 따라 유도된 수치값 I 를 계산한다.

$$I = (1/p - 1) \qquad (6\text{-}26)$$

D_M과 G 의 값은 식(6-27)에 의해 변환 후의 실험 데이터로부터 도출될 수 있다.

$$I = \exp(D - D_M) \cdot G \qquad (6\text{-}27)$$

이는 I 에 대한 로그 스케일을 사용하여 식(6-28)에 의해 선형 관계를 산출한다.

$$\log_e I = (D - D_M) \cdot G \qquad (6\text{-}28)$$

직선에 의한 보간은 간단하며, 경우에 따라서는 정확도가 충분하여 측정된 왜곡처도 D 에 의한 손상을 표현하는데 직선이 적절하다고 간주될 수 있다. 이 특성의 기울기는 식(6-29)와 같이 정의된다.

$$S = \frac{D_M - D}{\log_e I} = \frac{1}{G} \qquad (6\text{-}29)$$

이로부터 최적값 G 가 산출된다. 여기서, DM 은 I = 1 일 때의 D 값이다.

이 직선은 고려 중인 특정 손상과 관련된 손상 특성선이라고 할 수 있다. 또한, 이 직선은 로지스틱 함수의 특성값인 DM 과 G 에 의해 정의될 수 있음을 알 수 있다.

6.7.3.2 비대칭 함수에 의한 근사

6.7.3.2.1 함수 설명

실험점수와 영상 왜곡의 객관적 측정값 간의 관계를 대칭형 로지스틱 함수로 근사하는 방법은, 은 왜곡 파라미터 D 가 S/N (dB)과 같은 로그 단위로 측정되는 경우에 대부분 성공적이다. 그러나, 만약 왜곡 파라미터가 시간 지연(ms) 등과 같은 물리적 단위 d 로 측정된 경우에는, 식(6-27)을 식(6-30)으로 대체해야 한다.

$$I = (d/d_M)^{1/G} \qquad (6\text{-}30)$$

또한 식(6-25)는 식(6-31)과 같이 정의된다.

$$p = 1/\left[1 + (d/d_M)^{1/G}\right] \qquad (6\text{-}31)$$

이 함수는 비대칭적인 방식으로 로지스틱 함수를 근사한다.

6.7.3.2.2 근사 파라미터의 추정

실제 데이터와 함수 간의 잔여 오차를 최소화하는 함수의 최적 파라미터 추정은 임의의 재귀적 추정 알고리즘을 사용하여 구할 수 있다. 그림 6-3 은 비대칭 로지스틱 함수를 이용하여 실제 주관적 데이터를 근사한 예시를 나타낸다. 이러한 표현을 통해, 예를 들어 5 단계 척도에서 4.5 점에 해당하는 주관적 품질값에 대응하는 객관적 품질 지표를 추정할 수 있다.

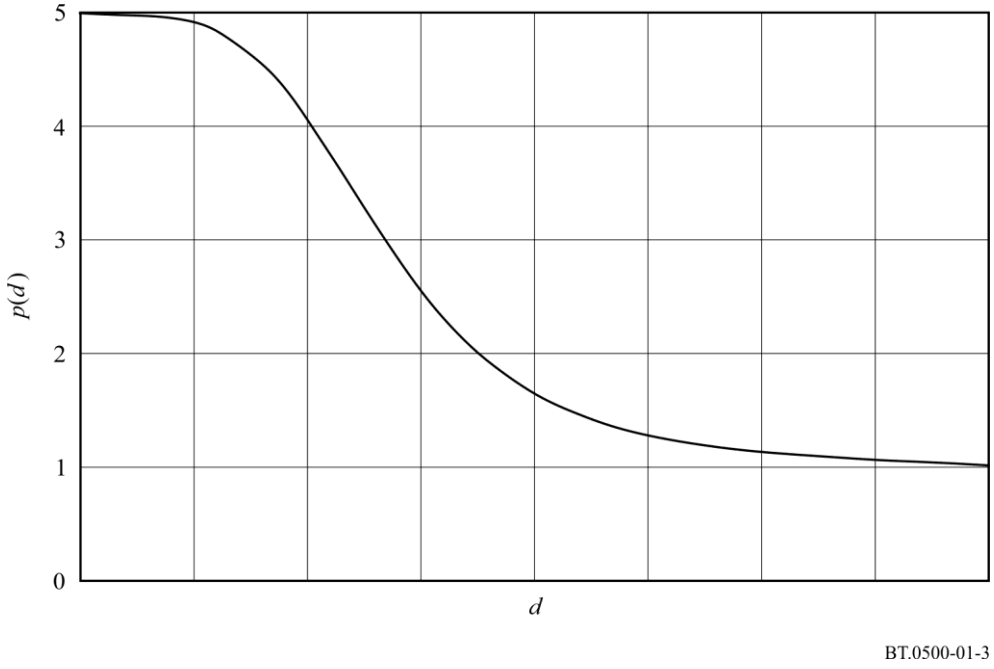


그림 6-3. 비대칭 추정

Non-symmetrical approximation

6.7.3.3. 잔여 손상/개선 및 척도 경계 효과의 보정

실제 실험에서 로지스틱 함수를 적용할 경우, 실험 데이터와 근사 결과 사이에 일부 차이가 발생하는 것을 피하기 어려운 경우가 있다. 이러한 불일치는 주로 척도의 양 끝단에서 발생하는 경계 효과 또는 여러 종류의 손상이 동시에 존재하여 통계 모델을 왜곡시키는 경우에 기인할 수 있다. 특히 척도 경계 효과라고 불리는 현상이

관찰되었는데, 평가자들이 평가척도의 극단값을 잘 사용하지 않는 경향을 보이는 것이다. 특히 고품질 영상일수록 심리적인 이유로 최고점을 주는 것을 주저하는 경향이 있다.

또한 식(6-1)에 따른 단순 산술평균을 척도 양 끝단 근처에서 사용할 경우, 점수 분포가 비정규(non-Gaussian) 형태를 보이기 때문에 평균값이 편향될 수 있다. 이러한 이유로, 종종 참조영상(기준영상) 조차도 평균점수(MOS)가 최고값(예: 5 점 만점 중 5 점)에 도달하지 못하는 잔여 손상이 보고되기도 한다. 따라서, 실험 데이터를 올바르게 해석하기 위해서는 경계 효과를 보정하는 것이 매우 중요하며, 여러 가지 유용한 보정 방법들이 존재한다(표 6-3 참조). 다만, 이러한 보정 절차들은 특정한 통계적 가정을 수반하기 때문에 매우 신중하게 적용해야 하며, 보정을 수행한 경우에는 그 내용을 반드시 결과 보고서에 명시해야 한다.

표 6-3. 척도 경계 효과 보정 기법의 비교
Comparison of methods of correction of the scale boundary effects

척도 경계 효과 보정 기법	특징		
무보정	잔여 손상 보정	잔여 개선 보정	척도 중심 이동 보정
	X	X	X
선형 척도 변환	O	유의미한 오차 발생 가능성 있음	X
비선형 척도 변환[4]	O	O	X
손상 합산 기반 보정 방법	O	X	O
곱셈 기반 보정 방법	O	X	O

[4] 비선형 척도 변환 방식을 적용하여, 보정된 평가 점수는 다음과 같이 계산한다.

6.7.3.4 그래프에 신뢰도 측면 통합

각 손상 조건에 대한 평균 등급과 이에 대응하는 95% 신뢰구간을 이용하여 다음의 세 가지 등급 시리즈를 구성한다:

- 최소 등급 시리즈: 평균값 – 신뢰구간
- 평균 등급 시리즈: 평균값 자체
- 최대 등급 시리즈: 평균값 + 신뢰구간

이 세 시리즈 각각에 대해 추정 파라미터(예: 로지스틱 함수의 계수)를 서로 독립적으로 계산한다. 이후, 세 함수의 결과를 그림 6-4 에 나타낸 바와 같이 하나의 그래프에 동시에 표시한다. 평균 시리즈는 실선으로, 최대/최소 시리즈는 점선으로 표시하며, 실험으로 얻은 실제 데이터 포인트(점)도 함께 표시한다. 이렇게 하면 연속적인 95% 신뢰 영역이 시각적으로 형성된다. 평가 척도상 4.5 점(가시성 한계점)에 해당하는 구간에서 그래프를 통해 직접 95% 신뢰구간을 읽어낼 수 있으며, 이 구간은 허용 오차 범위로 활용될 수 있다. 최대곡선과 최소곡선 사이의 영역은 진정한 95% 신뢰구간이 아니라, 평균적 근사치임을 유의해야 한다. 실험으로 얻은 데이터의 최소 95% 이상이 이 신뢰영역 안에 포함되어야 한다. 그렇지 않다면, 시험 수행 절차에 문제가 있었거나, 사용한 함수 모델(예: 로지스틱 근사)이 적절하지 않았음을 의미한다.

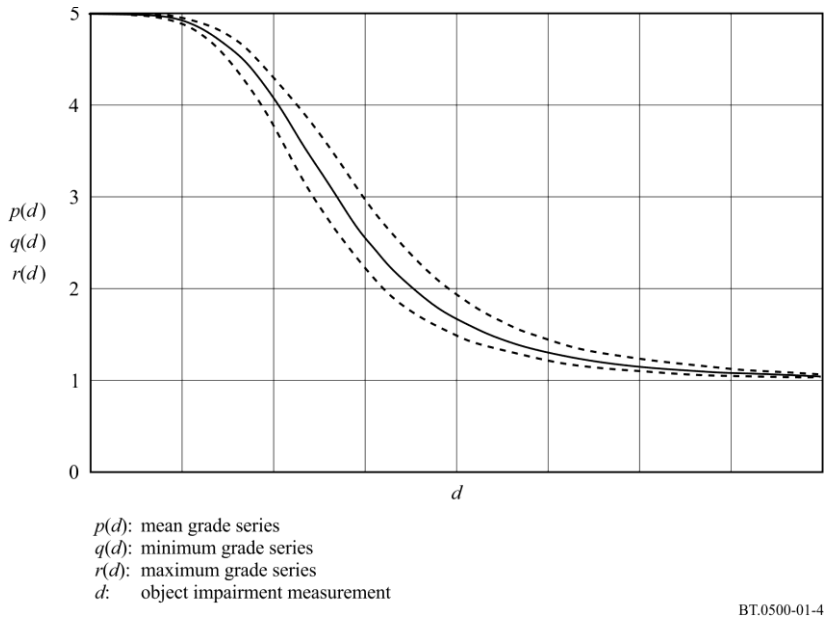


그림 6-4. 비대칭형 손상 특성의 사례

Case of non-symmetrical impairment characteristic

6.7.4 결과분석의 활용

신뢰구간, 즉 일련의 주관적 평가 시험의 정확도를 평가하기 위한 절차를 기술하였다. 이 절차는 고려 중인 특정 실험뿐만 아니라 동일한 방법론으로 수행된 다른 실험에도 관련된 평균적인 일반량을 추정하는 데에도 사용된다. 따라서 이러한 양들은 주관적 평가와 향후 실험 계획에 모두 도움이 되는 신뢰구간의 동작 다이어그램을 작성하는 데 사용될 수 있다.

7 주관적 화질 평가 방법론

이 파트는 주관적 영상 품질 평가를 수행하는 데 필요한 각 영상 평가 방법론의 세부 사항을 제공한다. 어떤 경우에는 6 장에 제시된 공통 평가 특징과 다를 수 있다.

다른 실험실에서 주관적 영상 품질 평가 결과를 올바르게 해석할 수 있도록 보장하려면, 절차에 대한 상세한 기록을 제공해야 하며 사용된 방법론의 모든 변형 사항은 평가 절차를 반복하고자 하는 다른 실험실에서 요구할 수 있는 모든 추가 정보와 함께 기록하는 것이 중요하다.

앞으로 설명할 모든 방법은 장점과 한계를 가지고 있으므로, 아직 다른 방법보다 어느 하나를 확실하게 추천하기는 불가능하다. 따라서 당면한 상황에 가장 적절한 방법을 선택하는 것은 연구자의 재량에 달려 있다.

다양한 방법들의 한계는 단일 방법에 너무 큰 비중을 두는 것이 현명하지 않을 수 있음을 시사한다. 따라서 여러 방법을 사용하거나 다차원적 접근법을 사용하는 것과 같은 보다 '완전한' 접근법을 고려하는 것이 적절할 수 있다.

7.1 DSIS(The double-stimulus impairment scale method)

7.1.1 일반 설명

일반적인 평가는 새로운 시스템 또는 전송 경로 손상의 효과에 대한 평가를 요구할 수 있다. 테스트 주최자의 초기 단계는 의미 있는 평가가 가능하도록 충분한 테스트 자료를 선택하고, 어떤 테스트 조건을 사용할지 결정하는 것을 포함한다. 만약 매개변수 변화의 효과가 중요하다면, 대략 동일한 간격의 적은 수의 단계로 손상 등급 범위를 포괄하는 매개변수 값 세트를 선택해야 한다. 매개변수 값을 변경할 수 없는 새로운 시스템이 평가되는 경우, 주관적으로 유사한 추가적인 손상을 더하거나, 또는 7.2 절의 DSCQS 와 같은 다른 방법을 사용해야 한다.

DSIS 방법(EBU 방법)은 평가자에게 먼저 손상되지 않은 참조 영상이 제시된 후 동일한 영상의 손상된 버전이 제시된다는 점에서 순환적이다. 그 후 평가자는 첫 번째 영상을 염두에 두고 두 번째 영상에 대해 평가하도록 해야 한다. 최대 30 분 동안 진행되는 세션에서, 평가자에게는 모든 필요한 조합을 포함하는 무작위 손상을 가진 일련의 영상 또는 시퀀스가 무작위 순서로 제시된다. 손상되지 않은 영상은 평가될 영상

또는 시퀀스에 포함된다. 일련의 세션이 끝나면 각 테스트 조건과 테스트 영상에 대한 평균 점수가 계산된다.

이 방법은 손상 척도를 사용하며, 이 척도에서는 일반적으로 결과의 안정성이 큰 손상보다 작은 손상에서 더 큰 것으로 나타난다. 이 방법이 때때로 제한된 범위의 손상과 함께 사용되기도 했지만, 전체 범위의 손상과 함께 사용하는 것이 더 적절하다.

7.1.2 일반 구성

시청 조건, 소스 신호, 테스트 자료, 관찰자 및 결과 발표 방식은 6 장에 따라 정의되거나 선택된다.

테스트 시스템의 일반적인 구성은 그림 7-1 과 같아야 한다.

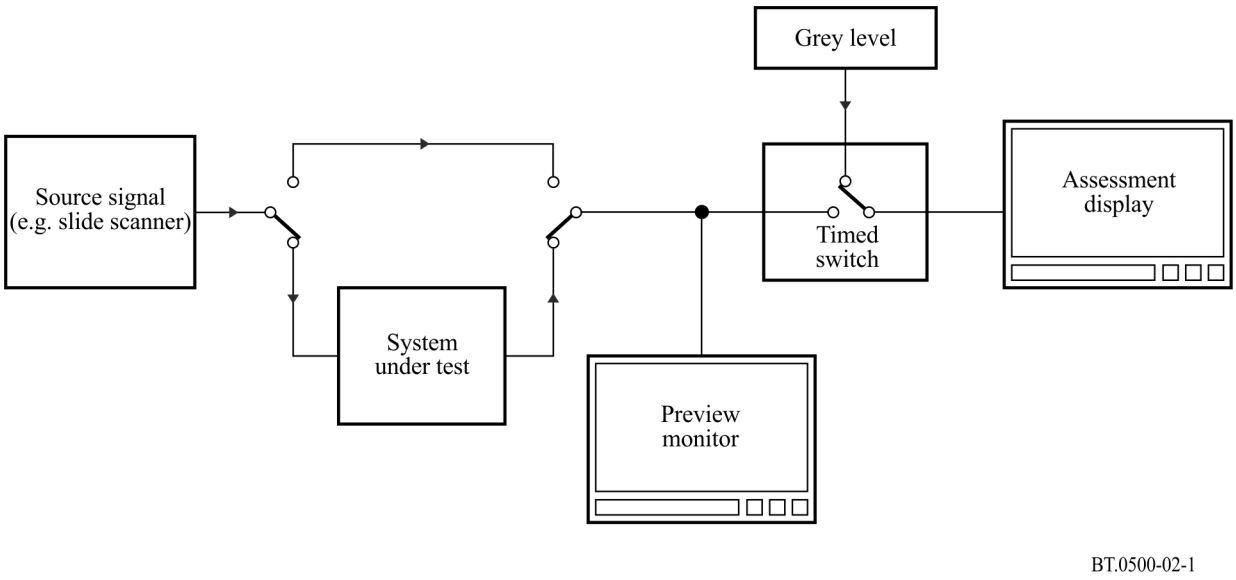


그림 7-1. DSIS 테스트 시스템의 기본 구조

GENERAL ARRANGEMENT FOR TEST SYSTEM FOR DSIS METHOD

평가자는 시간 스위치를 통해 신호가 공급되는 평가용 디스플레이를 본다. 시간 스위치로 가는 신호 경로는 소스 신호에서 직접 오거나, 테스트 중인 시스템을 통해 간접적으로 올 수 있다. 평가자에게는 일련의 테스트 영상 또는 시퀀스가 제시된다. 이것들은 쌍으로 배열되는데, 쌍의 첫 번째는 소스에서 직접 오고, 두 번째는 테스트 중인 시스템을 통한 동일한 영상이다.

7.1.3 테스트 자료 제시

테스트 세션은 여러 번의 제시로 구성된다. 제시 구조에는 아래에 설명된 I 과 II 의 두 가지 변형이 있다.

변형 I: 참조 영상 또는 시퀀스와 테스트 영상 또는 시퀀스가 그림 2-2(a)와 같이 한 번만 제시된다.

변형 II: 참조 영상 또는 시퀀스와 테스트 영상 또는 시퀀스가 그림 2-2(b)와 같이 두 번 제시된다.

변형 II: 변형 I 보다 시간이 더 많이 소요되며, 매우 작은 손상의 구별이 필요하거나 동영상 시퀀스를 테스트하는 경우에 적용될 수 있다.

7.1.4 등급 척도

5 등급 손상 척도를 사용해야 한다:

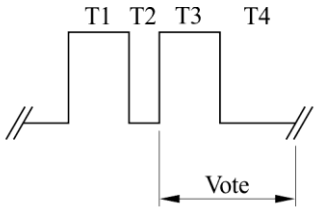
5 감지할 수 없음 (imperceptible)
4 감지할 수 있으나, 거슬리지 않음 (perceptible, but not annoying)
3 약간 거슬림 (slightly annoying)
2 거슬림 (annoying)
1 매우 거슬림 (very annoying).

평가자는 척도를 매우 명확하게 보여주고, 등급을 기록하기 위한 번호가 매겨진 상자나 다른 수단이 있는 양식을 사용해야 한다.

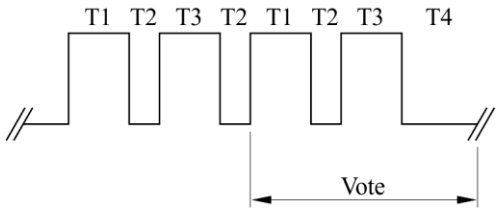
7.1.5 평가에 대한 소개

각 세션이 시작될 때, 관찰자에게 평가 유형, 등급 척도, 순서 및 시간(참조 영상, 회색 화면, 테스트 영상, 투표 시간)에 대한 설명이 제공된다. 평가될 손상의 범위와 유형은 테스트에 사용된 영상이 아닌, 비슷한 민감도를 가진 다른 영상을 통해 예시되어야 한다. 보이는 최악의 품질이 반드시 가장 낮은 주관적 등급에 해당한다고 암시해서는 안 된다. 관찰자는 영상이 주는 전반적인 인상에 근거하여 판단하고, 이러한 판단을 주관적 척도를 정의하는 데 사용된 용어로 표현해야 한다.

관찰자는 T1 과 T3 시간 동안 내내 영상을 보아야 한다. 투표(평가)는 T4 동안에만 허용되어야 한다.



a) Variant I



b) Variant II

BT.0500-02-2

그림 7-2. 테스트 자료의 표현 방식

PRESENTATION STRUCTURE OF TEST MATERIAL

제시 단계

T1 = 10 s	참조 영상 (Reference image)
T2 = 3 s	약 200 mV 비디오 레벨로 생성된 중간 회색 (Mid-grey)
T3 = 10 s	시험 조건(Test condition)
T4 = 5 to 11 s	중간 회색 (Mid-grey)

경험적으로 볼 때, T1과 T3 구간을 10초 이상으로 늘려도 평가자가 영상 시퀀스를 등급화(평가)하는 능력이 향상되지는 않는다.

7.1.6 테스트 세션

영상과 손상은 의사-무작위(pseudo-random) 순서로 제시되어야 하며, 가급적 각 세션마다 다른 순서로 제시되어야 한다. 어떠한 경우에도, 손상 수준이 같거나 다르더라도 동일한 테스트 영상 또는 시퀀스가 연속으로 두 번 제시되어서는 안 된다.

손상의 범위는 대부분의 관찰자가 모든 등급을 사용하도록 선택되어야 하며, 3 에 가까운 총 평균 점수(실험에서 이루어진 전체 평가의 평균)를 목표로 해야 한다.

한 세션은 설명과 예비 절차를 포함하여 대략 30 분을 넘지 않아야 한다. 테스트 시퀀스는 손상의 범위를 보여주는 몇 개의 영상으로 시작할 수 있으며, 이 영상들에 대한 평가는 최종 결과에 포함되지 않는다.

7.2 DSCQS(The double-stimulus continuous quality-scale method)

7.2.1 일반 설명

일반적인 평가의 목적은 새로운 시스템의 화질을 평가하거나, 전송 경로가 화질에 미치는 영향을 평가하는 것이다. DS 방법은 테스트 자극 조건에서 화질의 전 범위를 충분히 재현하기 어려운 경우에 특히 유용하다.

이 방법은 순환적 특성을 가지는데, 평가자는 동일한 원본에서 얻은 두 개의 영상을 순서대로 시청하게 된다. 그중 하나는 시험 대상 프로세스를 거친 것이고, 다른 하나는 원본 영상 그대로이다. 평가자는 두 영상의 화질을 모두 평가한다.

평가 세션은 최대 30 분 동안 진행되며, 이 동안 평가자에게 여러 쌍의 영상이 제시된다. 이 영상 쌍들은 내부적으로 무작위 순서로 제시되며, 모든 필요한 조합을 포함하는 다양한 열화가 임의로 적용된다. 세션 종료 후, 각 테스트 조건과 테스트 영상에 대한 평균 점수가 계산된다.

7.2.2 일반 구성

시청 조건, 소스 신호, 테스트 자료, 관찰자 및 평가 소개 방식은 6 장과 7 장에 따라 정의되거나 선택된다. 테스트 세션은 7.1.6 절에 설명된 절차를 따른다. 테스트 시스템의 일반적인 구성은 그림 7-3 과 같이 제시되어야 한다.

7.2.3 테스트 자료 제시

하나의 테스트 세션은 여러 번의 프레젠테이션으로 구성된다.

변형 I 의 경우, 단일 관찰자가 참여하며, 각 프레젠테이션에서 평가자는 A 신호와 B 신호 사이를 자유롭게 전환할 수 있다. 이 과정은 평가자가 각 신호에 대한 품질의 심리적 기준을 형성할 때까지 반복될 수 있다. 평가자는 보통 이러한 전환을 2~3 회 정도, 각각 최대 약 10 초 이내의 시간 동안 수행한다.

변형 II 는 다수의 관찰자가 동시에 참여하는 방식이다. 결과를 기록하기에 앞서, 평가자가 각 조건의 품질에 대한 심리적 기준을 형성할 수 있도록 조건 쌍을 동일한 시간동안 한 번 이상 제시한다. 그 후, 동일한 조건 쌍을 한 번 이상 다시 제시하면서 평가 결과를 기록한다. 반복 횟수는 테스트 시퀀스(시험 영상)의 길이에 따라 달라진다. 정지 영상의 경우, 3~4 초 길이의 시퀀스를 5 회 반복하여(마지막 두 번 동안 평가를 수행하는 방식으로) 제시하는 것이 적절하다. 시간적으로 변화하는 영상 열화를 포함한 동영상의 경우에는, 약 10 초 길이의 시퀀스를 두 번 반복(두 번째 제시에서 평가를 수행)하는 것이 적절하다. 프레젠테이션의 구조는 그림 7-4 에 제시되어 있다.

현실적인 제약으로 인해 사용 가능한 시퀀스의 길이가 10 초 미만인 경우, 이러한 짧은 시퀀스를 세그먼트로 구성하여 전체 표시 시간을 10 초로 확장할 수 있다. 이때, 세그먼트 간의 불연속성을 최소화하기 위해 연속된 시퀀스 세그먼트를 시간 순서 반대로 하여 구성할 수 있다. 이러한 방식을 '팔린드로믹 표시(palindromic display)'라고 부른다. 역방향 시간 세그먼트를 사용할 때는 주의해야 한다. 테스트 조건이 인과적 과정을 올바르게 표현하도록 반전된 시퀀스는 반드시 역방향 시간으로 변환된 원 신호를 시험 대상 시스템에 통과시켜 얻어야 한다.

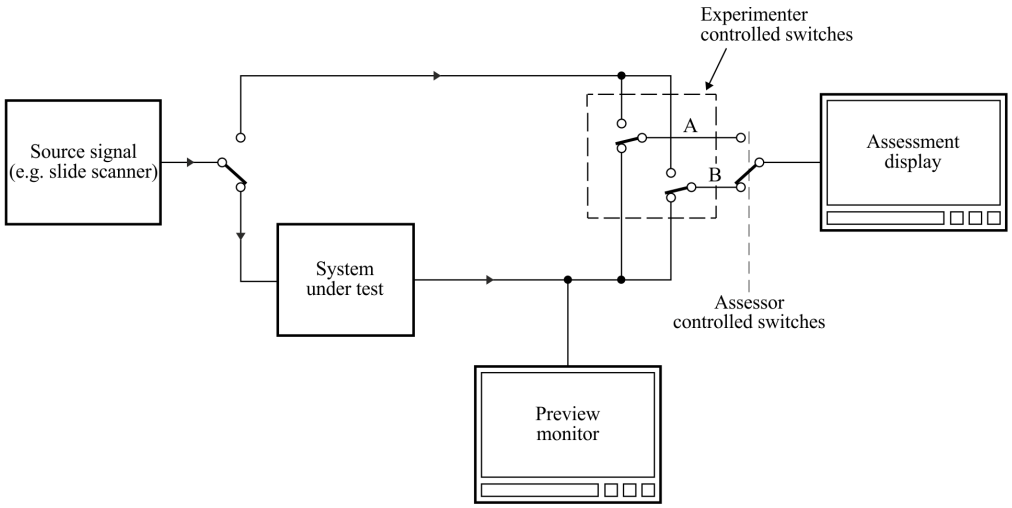


그림 7-3. DSCQS 테스트 시스템의 기본 구조

GENERAL ARRANGEMENT FOR TEST SYSTEM FOR DSCQS METHOD

이 방법에는 아래에 설명된 두 가지 변형이 있다.

- **변형 I**

평가자는 보통 단독으로 시험에 참여하며, A 조건과, B 조건 사이를 자유롭게 전환하면서 각 조건의 화질에 대한 명확한 판단이 설립될 때까지 관찰할 수 있다. A 라인과 B 라인에는 참조 영상 또는 테스트 중인 시스템을 거친 영상이 공급되지만, 어느 영상이 어느 라인으로 공급되는지는 한 테스트 조건에서 다음 테스트 조건으로 넘어갈 때 무작위로 변경된다. 이 조건은 실험자에게는 알려지지만 평가자에게는 알려지지 않는다.

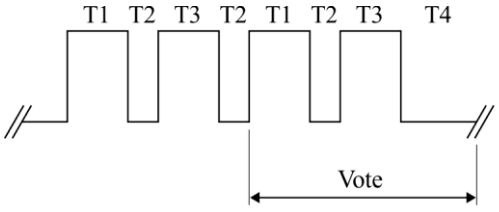
- **변형 II**

여러 명의 평가자에게 A 라인과 B 라인으로부터의 영상이 순차적으로 제시된다. 이를 통해 각 조건의 화질에 대한 판단을 형성한다. 각 프레젠테이션에서 A 라인과 B 라인의 영상 입력은 변형 I과 동일한 방식으로 구성된다.

7.2.4 등급 척도

이 방법에서는 각 시험 영상에 대해 두 가지 버전을 평가해야 한다. 각 영상 쌍 중 하나는 열화되지 않은 영상이며, 다른 하나는 열화를 포함할 수도 있고 포함하지 않을 수도 있다. 열화되지 않은 영상은 참조 용으로 포함되지만, 관찰자에게는 어느 영상이 참조 영상인지는 알려주지 않는다. 일련의 테스트에서 참조 영상의 위치는 의사-무작위(pseudo-random) 순서로 변경된다.

관찰자는 단순히 각 영상의 전반적인 화질을 평가하도록 요청되면, 이를 위해 수직 척도에 표시하도록 한다. 수직 척도는 각 시험 영상의 이중 제시를 수용할 수 있도록 쌍으로 인쇄되어 있다. 이 척도는 연속 등급 체계를 제공하여 양자화 오류를 피할 수 있도록 하였으며, 실제 인쇄 시에는 ITU-R 의 표준 5 단계 화질 척도에 대응하도록 다섯 구간으로 등분되어 있다. 각 등급에 대응하는 용어는 기존 ITU-R 에서 일반적으로 사용하는 등급 명칭과 동일하지만, 여기서는 참고용으로만 포함되면, 점수 기록지 상 10 개의 이중 열 중 각 행의 첫 번째 열의 왼쪽에만 인쇄된다. 그림 7-5 는 일반적인 점수 기록지의 일부를 보여준다. 척도 구분선과 실제 평가 결과가 혼동되지 않도록, 척도는 청색으로 인쇄하고, 평가 결과는 흑색으로 기록하도록 한다.



BT.0500-02-4

그림 7-4. 테스트 자료의 표현 방식

PRESENTATION STRUCTURE OF TEST MATERIAL

제시 단계	
T1 = 10 s	참조 영상(Reference image)
T2 = 3 s	약 200 mV 비디오 레벨로 생성된 중간 회색(Mid-grey)
T3 = 10 s	시험 조건(Test condition)
T4 = 5 to 11 s	중간 회색(Mid-grey)

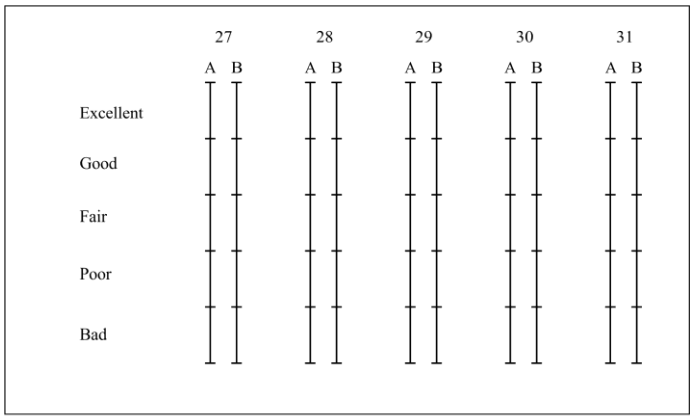


그림 7-5. 연속 척도를 사용한 화질 평가 양식의 일부5

PORTION OF QUALITY-RATING FORM USING CONTINUOUS SCALES

7.2.5 결과 분석

각 테스트 조건에 대해 얻어진 평가 쌍(즉, 기준 영상과 테스트 영상)은 점수 표 상의 길이 측정값을 0 에서 100 사이의 정규화 된 점수로 변환한다. 그 후 참조 조건과 테스트 조건 간의 평가 점수 차이를 계산한다.

경험적으로, 서로 다른 테스트 시퀀스에 대해 얻은 점수는 사용된 테스트 자료의 중요도에 따라 달라진다. 평가에 사용된 모든 테스트 시퀀스의 결과를 단순히 평균한 값으로 제시하기보다 다른 테스트 시퀀스에 대한 결과를 개별적으로 제시함으로써 코덱 성능에 대한 더 완전한 이해를 얻을 수 있다.

개별 테스트 시퀀스에 대한 결과를 테스트 시퀀스 중요도의 순위 순서로 정렬하여 가로축에 배열하면, 테스트 중인 시스템의 영상 콘텐츠 열화 특성에 대해 대략적으로 시각화한 그래프 형태로 표시할 수 있다. 그러나 이러한 제시 형태는 코덱의 성능 특성만을 설명할 뿐, 특정 레벨의 중요도를 가진 시퀀스가 실제로 발생할 가능성은 보여주지 못한다. 시스템 성능을 보다 완전하게 이해하기 위해서는 테스트 시퀀스의

---

[5] DSCQS 방법을 위한 테스트 세션 내 테스트 항목의 배치를 계획할 때, 실험에 체계적 오류가 없음을 확인할 수 있도록 검증 절차를 포함하는 것이 바람직하다. 그러나 이러한 신뢰도 검증의 구체적인 수행 방법은 현재 연구 및 검토가 진행 중이다.

중요도와 특정 레벨의 중요도를 가진 시퀀스가 발생할 확률에 대한 추가 연구가 필요하다.

7.2.6 결과 해석

DSCQS 을 사용할 때, 다른 시험 절차에서 사용하는 형용사적 등급(adjectival terms)-예를 들어 DSIS 에서 사용되는 용어인 '감지 불가', 감지 가능하나 거슬리지 않음' 등-을 DSCQS 점수에 직접 연관시켜 테스트 조건의 화질 레벨을 해석하는 위험하며, 잘못된 결론을 초래할 수 있다.

DSCQS 에서 얻은 결과는 절대 점수가 아니라 참조 조건과 테스트 조건 간의 점수 차이로 다루어져야 한다는 점에 유의해야 한다. 따라서 DSCQS 프로토콜 자체에서 온 용어(예: 우수 Excellent, 좋음 Good, 보통 Fair)라 할지라도 점수를 단일 품질 설명 용어와 연관시키는 것은 잘못된 것이다.

어떤 테스트 절차에서든 평가를 시작하기 전에 수용 기준을 미리 결정하는 것이 중요하다. 이는 경험이 없는 사용자가 이 방법으로 얻은 화질 척도 값의 의미를 해석하는 경향이 있기 때문에 DSCQS 을 사용할 때 특히 중요하다.

7.3 SS(Single-stimulus method)

SS 에서는 단일 영상 또는 영상 시퀀스가 제시되고, 평가자는 전체 제시물에 대한 종합적 평가 지표를 제공한다. 테스트 자료는 테스트 시퀀스 만을 포함할 수도 있고, 테스트 시퀀스와 그에 대응하는 참조 시퀀스를 모두 포함할 수도 있다. 참조 시퀀스를 포함하는 후자의 경우, 참조 시퀀스는 다른 테스트 자극과 동일하게 독립된 평가 자극으로 제시된다.

7.3.1 일반 배치

시청 조건, 소스 신호, 조건의 범위 및 앵커링(anchoring), 관찰자, 평가 절차의 소개 및 결과 제시 방식은 6 장에 명시된 기준에 따라 정의되거나 선택되어야 한다.

7.3.2 테스트 자료 선택

실험실 테스트의 경우, 테스트 영상의 내용은 6.3 절에서 기술된 기준에 따라 선정해야 한다.

내용이 선정되면, 테스트 영상은 검토 중인 설계 변수 또는 하나 이상의 요인의 범위를 반영하도록 제작한다. 두 개 이상의 요인을 동시에 평가할 경우, 영상은 다음의 두 가지 방식 중 하나로 준비할 수 있다.

첫 번째 방법: 각 영상은 하나의 요인에 의한 한 레벨 만을 나타낸다.

두 번째 방법: 각 영상은 모든 요인에 의한 한 레벨을 포함하되, 전체 영상 집합에서는 각 요인의 모든 레벨이 다른 요인의 레벨들과 조합되도록 구성한다.

시험 영상을 만들 때, 시험 요인을 체계적으로 바꿔가며 구성하면 결과가 어떤 요인 때문에 달라졌는지를 명확하게 구분할 수 있다.

예)요인 1: 압축률(low/high), 요인 2: 해상도(HD/UHD)

영상 조합을 잘 설계하면 “화질이 떨어진 이유가 압축률 때문인지, 해상도 때문인지” 통계적으로 분리해서 알 수 있다.

두 가지 방법 모두 결과가 어떤 요인에 의해 나타났는지를 명확하게 구분할 수 있게 해주며, 두 번째 방법은 두 요인이 단순히 각각의 영향만 미치는 것이 아니라, 서로 결합할 때 다른 영향을 만드는 경우 요인간 상호 작용으로, 복합적 영향을 탐지할 수도 있다.

7.3.3 테스트 세션

테스트 세션은 일련의 평가 시행으로 구성된다. 이들은 무작위 순서(random order)로 제시되어야 하며, 가급적 관찰자마다 서로 다른 무작위 순서로 제시되는 것이 바람직하다. 하나의 동일한 무작위 순서를 사용할 경우 다음의 두가지 제시 구조 변형[I SS 방식 및 II SSMR 방식]이 있다.

a) 변형 I (SS 방식)

테스트 영상 또는 시퀀스는 테스트 세션에서 한 번만 제시된다. 첫 세션의 시작 부분에는 몇 개의 더미 시퀀스가 도입되어야 한다 (6.7 절에 설명된 대로).

일반적으로 실험은 동일한 레벨의 열화를 가진 동일 영상이 연속하여 두 번 제시되지 않도록 설계된다.

일반적인 평가 시도는 세 개의 화면으로 구성된다.

- 중간 회색 적응 화면(mid-grey adaptation field)
- 자극(stimulus)
- 중간 회색 후노출 화면(mid-grey post-exposure field)

이 세 화면의 표시 시간은 시청자 과제(viewer task), 시험자료(material) 및 평가 요인에 따라 달라질 수 있으나, 보통 각각 3 초, 10 초, 10 초로 설정하는 것이 일반적이다. 평가 지표(viewer index)는 자극 또는 후노출 필드 표시 중에 수집될 수 있다.

b) 변형 II (SSMR 방식)

테스트 영상 또는 시퀀스는 세 번 제시되며, 이에 따라 테스트 세션은 세 개의 프레젠테이션으로 구성된다. 각 프레젠테이션에는 테스트할 모든 영상 또는 시퀀스가 한 번씩만 포함된다. 각 프레젠테이션의 시작은 화면에 “Presentation 1 “과 같은 메시지를 표시하여 시작을 알린다. 첫 번째 프레젠테이션은 관찰자의 의견을 안정화하는데 사용되며, 이 프레젠테이션에서 얻은 데이터는 테스트 결과에 포함되지 않는다. 각 영상 또는 시퀀스에 부여된 최종 점수는 두번째와 세번째 프레젠테이션에서 얻은 데이터의 평균값으로 계산된다. 실험에서는 일반적으로 각 프레젠테이션 내의 영상 또는 시퀀스의 무작위 순서에 대해 다음의 제약 조건이 적용되도록 한다.

- 주어진 영상 또는 시퀀스가 다른 프레젠테이션에서 동일한 위치에 있지 않아야 함
- 주어진 영상 또는 시퀀스가 다른 프레젠테이션에서 동일한 영상 또는 시퀀스 바로 앞에 위치하지 않아야 함

일반적인 평가 시행은 자극과 중간 회색 후노출 화면의 두 가지 디스플레이로 구성된다. 이러한 디스플레이의 지속 시간은 시청자 과제, 시험 자료 및 평가되는 의견이나 요인에 따라 다를 수 있지만, 각각 10 초와 5 초가 권장된다. 시청자 지표는 후노출 필드 표시 중에만 수집되어야 한다.

변형 II(SSMR)는 각 시험 영상 또는 시퀀스를 평가하는데 소요되는 시간이 약 45 초대 23 초로, 변형 I 에 비해 명확히 더 길다는 단점이 있다. 그러나 이 방법은 변형 I 에서 나타나는 영상이나 시퀀스의 제시 순서에 따른 결과의 편향 의존성을 줄여준다. 또한 실험 결과에 따르면 변형 II 는 투표 결과의 범위 내에서 약 20% 정도의 분산을 허용하는 것으로 나타났다.

7.3.4 SS 방법의 유형

일반적으로, 텔레비전 화질 평가에서는 세 가지 유형의 단일 자극 방법이 사용되어 왔다.

7.3.4.1 형용사적 범주 평가 방법

형용사적 범주 평가에서, 관찰자는 영상 또는 영상 시퀀스를 일반적으로 의미론적 용어로 정의되는 범주 집합 중 하나에 할당한다. 이러한 범주는 특정 속성이 감지되었는지 여부에 대한 판단(예: 손상 임계값을 설정하기 위한 경우)을 반영할 수 있다. 영상 품질과 영상 열화를 평가하는 범주형 척도가 가장 자주 사용되어 왔으며, ITU-R 에서 권고하는 척도는 표 7-1 에 나와 있다. 운영 환경에서는 때때로 절반 등급이 사용되는 경우도 있다. 텍스트 가독성, 읽기 노력, 영상 유용성을 평가하는 척도는 특별한 경우에 사용되었다.

표 7-1. ITU-R 품질 및 결함 등급 기준
ITU-R quality and impairment scales

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

이 방법은 각 테스트 조건에 대해 척도 범주에 걸친 판정 결과 분포를 산출한다. 응답을 분석하는 방식은 판단(탐지 등)과 구하고자 하는 정보(탐지 임계값, 조건의 순위 또는 중심 경향, 조건 간의 심리적 "거리")에 따라 다르다. 많은 분석 방법이 사용 가능하다.

7.3.4.2 수치적 범주 평가 방법

11 단계 수치 범주 척도(SSNCS)를 사용하는 SS 절차가 그래픽 척도 및 비율 척도와 비교하여 연구되었다. ITU-R BT.1082 보고서에 기술된 이 연구는 SSNCS 방법이 민감도와 안정성 측면에서 명확하게 우수하다는 것을 보여준다.

7.3.4.3 비범주적 평가 방법

비범주적 판단에서 관찰자는 제시된 각 영상 또는 영상 시퀀스에 값을 할당한다. 이 방법에는 두 가지 형태가 있다.

1) 연속 척도법(선 위의 위치 표시)

연속 척도법은 범주 판단의 한 변형으로, 평가자는 각 영상이나 시퀀스를 두 개의 의미론적 레이블(예: 표 7-1 의 범주형 척도의 양 끝)을 잇는 선 위의 한 지점에 표시한다. 이 척도는 참조를 위해 중간 지점에 추가 레이블을 포함할 수도 있다. 각 조건에 대한 지표(index)는 척도 한쪽 끝에서 해당 지점까지의 거리로 정의된다.

2) 수치 척도법(숫자 평가)

수치 척도법에서는 평가자가 각 영상 또는 영상 시퀀스에 특정 차원(예: 영상 선명도)에서 판정된 레벨을 반영하는 숫자를 할당한다. 사용되는 숫자의 범위는 제한될 수도(예: 0-100) 있고 그렇지 않을 수도 있다. 때로는 할당된 숫자가 (일부 형태의 크기 추정에서처럼 다른 영상이나 영상 시퀀스의 레벨을 직접 참조하지 않고) '절대적인' 기준에 따라 판정된 레벨을 나타내는 경우도 있다. 다른 경우에는 이전에 제시된 '표준 영상'에 대한 상대적인 기준으로 판정된 레벨을 나타낸다. (예: 크기 추정, 분할법, 비율 추정).

두 형태 모두 각 조건에 대해 수치 분포로 표현되며, 분석은 평가 유형과 목적에 따라 순위, 평균, 심리적 거리 등 다양한 통계적 방법을 사용할 수 있다.

7.3.4.4 수행 방법

정상적인 시청의 일부 측면은 외부 지향적 과제(목표 정보 찾기, 텍스트 읽기, 객체 식별 등)의 수행 능력으로 표현될 수 있다. 이러한 경우 그 과제가 수행되는 정확도나 속도와 같은 성능 지표가 영상 또는 영상 시퀀스의 품질을 나타내는 지표로 사용될 수 있다.

수행 방법은 각 시험 조건에 대한 정확도 또는 속도 점수의 분포를 생성한다. 이후 분석은 점수의 중심 경향과 분산을 비교하여 조건 간의 관계를 규명하는데 초점을 둔다. 이때 일반적으로 분산 분석 또는 이와 유사한 통계 기법이 사용된다.

7.4 SC(Stimulus-Comparison method)

SC 에서는 두 개의 영상 또는 영상 시퀀스가 표시되고, 시청자는 두 제시 간의 관계에 대한 지표를 제공한다.

7.4.1 일반 배치

시청 조건, 소스 신호, 조건 범위 및 앵커링, 관찰자, 평가 소개 및 결과 발표 방식은 6 장에 따라 정의되거나 선택된다.

7.4.2 테스트 자료 선택

사용되는 영상 또는 영상 시퀀스는 SS 와 동일한 방식으로 생성된다. 결과로 나온 영상 또는 영상 시퀀스는 평가 시행에 사용될 쌍을 형성하기 위해 결합된다.

7.4.3 테스트 세션

평가 시행은 하나의 디스플레이 또는 잘 일치된 두 개의 디스플레이를 사용하며, 일반적으로 SS 경우와 같이 진행된다. 만약 하나의 디스플레이가 사용되면, 시행은 첫 번째와 지속 시간이 동일한 추가적인 자극 필드를 포함한다. 이 경우, 시행 전반에 걸쳐

쌍의 두 구성원이 첫 번째와 두 번째 위치에 동일하게 자주 나타나도록 하는 것이 좋다. 만약 두 개의 디스플레이가 사용되면, 자극 필드는 동시에 표시된다.

SC 는 판단이 모든 가능한 조건 쌍을 비교할 때 조건 간의 관계를 더 정확히 평가한다. 그러나 이것이 너무 많은 수의 관찰을 요구하는 경우, 관찰을 평가자들 사이에 나누거나 모든 가능한 쌍의 샘플을 사용하는 것이 가능할 수 있다.

7.4.4 SS 유형

텔레비전 평가에서는 세 가지 유형의 SC 가 사용되어 왔다.

7.4.4.1 형용사적 범주 판단 방법

형용사적 범주 판단 방법에서, 관찰자는 한 쌍의 구성원 간의 관계를 일반적으로 의미론적 용어로 정의되는 범주 집합 중 하나에 할당된다. 이러한 범주는 인지 가능한 차이의 존재 여부(예: 같음, 다름), 인지 가능한 차이의 존재 및 방향(예: 더 나쁨, 같음, 더 좋음), 또는 정도와 방향에 대한 판단을 나타낼 수 있다. ITU-R 비교 척도는 표 7-2 에 나와 있다.

표 7-2. 비교 기준
Comparison scale

-3	Much worse
-2	Worse
-1	Slightly worse
0	The same
+1	Slightly better
+2	Better
+3	Much better

이 방법은 각 조건 쌍에 대해 척도 범주에 걸친 판단의 분포를 산출한다. 응답이 분석되는 방식은 내려진 판단(예: 차이)과 요구되는 정보(예: 최소 식별 차이(JND), 조건의 순위, 조건 간의 '거리' 등)에 따라 다르다.

7.4.4.2 비범주적 판단 방법

비범주적 판단에서, 관찰자는 평가 쌍의 구성원 간의 관계에 값을 할당한다. 이 방법에는 두 가지 형태가 있다:

- 연속 스케일링에서, 평가자는 각 관계를 두 레이블(예: 같음-다름 또는 표 2-2와 같은 범주형 척도의 양 끝) 사이에 그려진 선 위의 한 점에 할당한다. 척도는 중간 지점에 추가적인 참조 레이블을 포함할 수 있다. 선의 한쪽 끝에서의 거리가 각 조건 쌍의 값으로 간주된다.

- 두 번째 형태에서, 평가자는 각 관계에 특정 차원(예: 품질 차이)에서 판단된 수준을 반영하는 숫자를 할당한다. 사용되는 숫자의 범위는 제한될 수도 있고 그렇지 않을 수도 있다. 할당된 숫자는 '절대적인' 용어로 또는 '표준' 쌍에서의 관계에 대한 상대적인 용어로 관계를 설명할 수 있다.

두 형태 모두 각 조건 쌍에 대한 값의 분포가 결과로 도출된다. 분석 방법은 판단의 성격과 요구되는 정보에 따라 달라진다.

7.4.4.3 수행 방법

어떤 경우에는, 수행 척도가 자극-비교 절차에서 도출될 수 있다. 강제 선택법에서는 한 구성원은 특정 수준의 속성(예: 손상)을 포함하고 다른 구성원은 다른 수준 또는 속성 없음을 포함하도록 쌍이 준비된다. 관찰자는 어느 구성원이 더 크거나/작은 수준의 속성을 포함하는지 또는 어느 것이 속성을 포함하는지를 결정하도록 요청 받는다. 수행의 정확도와 속도는 쌍의 구성원 간의 관계에 대한 지표로 간주된다.

7.5 SSCQE(Single stimulus continuous quality evaluation)

디지털 텔레비전 압축의 도입은 장면 특성에 따라 달라지고 시간에 따라 변화하는 영상 품질 손상을 야기한다. 디지털로 코딩된 비디오의 짧은 구간에서도 장면의 복잡성,

움직임, 세부 묘사 정도에 따라 품질이 크게 변동할 수 있으며, 손상은 짧게 발생했다가 사라지는 형태로 나타난다. 이러한 특성 때문에 기존 ITU-R 에서 제시한 고정적이고 평균 기반의 평가 방법만으로는 충분하지 않다.

특히, 실험실 환경에서 활용되는 DS 기반 방법은 실제 방송이나 서비스 상황과는 다르게, 피험자가 원본 영상을 항상 참조할 수 있다는 가정을 전제로 한다. 이는 가정 내에서 실제 시청자가 경험하는 조건과 괴리가 존재한다. 따라서 피험자가 소스 참조 없이 자료를 한 번만 시청하면서 디지털로 코딩된 비디오의 주관적 품질을 지속적으로 평가하는 기법이 필요하다. 이러한 요구로 인해 SSCQE 가 개발되고 검증되었다.

7.5.1 기록 장치 및 설정

연속적인 품질 평가 데이터를 기록하기 위해, 피험자에게 전자식 기록 핸드셋을 제공해야 한다. 이 장치는 컴퓨터에 연결되어야 하며, 다음과 같은 요건을 충족해야 한다.

- 스프링 복귀점이 없는 슬라이더 메커니즘
- 10cm 의 선형 이동 범위
- 고정식 또는 책상 위 장착 가능 구조
- 초당 2 회 샘플링으로 데이터 기록

이러한 설정은 피험자가 품질 변화를 연속적으로 표현할 수 있게 하며, 순간적인 화질 열화나 회복을 시간 축에서 정밀하게 포착할 수 있게 한다.

7.5.2 테스트 프로토콜

테스트 절차는 다음과 같은 구조를 따른다.

- **프로그램 세그먼트 (PS)**
  - 특정 품질 매개변수(QP, 예: 비트레이트)에 따라 처리된 하나의 프로그램 유형(예: 스포츠, 뉴스, 드라마)으로 정의한다.
  - 각 PS 는 최소 5 분 이상 지속되어야 한다.
- **테스트 세션 (TS)**
  - 의사 무작위 순서로 배열된 PS/QP 조합의 연속으로 구성한다.
  - 각 TS 는 모든 PS 와 QP 를 최소 한 번 포함해야 하나, 모든 PS/QP 조합이 반드시 포함될 필요는 없다.
  - TS 의 길이는 30 분에서 60 분 사이여야 한다.

- 테스트 제시 (TP)
    - 테스트 전체를 의미한다.
    - PS/QP 조합의 수가 많을 경우 여러 TS 로 나눠 진행한다.
    - 조합 수가 제한적일 경우 동일한 TS 를 반복하여 충분한 지속 시간을 확보할 수 있다.

서비스 품질 평가를 위해 오디오를 포함할 수 있으며, 이 경우 오디오 콘텐츠의 선정 또한 비디오와 동일한 중요도로 고려해야 한다. 가장 단순한 테스트 형식은 단일 PS 와 단일 QP 를 사용하는 것이다.

7.5.3 시청 조건

시청 조건은 6 장에 명시된 일반 조건이나, 8 장을 따른다. 이는 실제 시청 환경과 유사한 조건을 보장하기 위해 필수적이다.

- 일반 시청 조건 (6 장 기준)
    - **실험실 환경**: 시스템을 엄격하게 검증하기 위해 조도는 낮게 유지하고, 배경의 색온도는 D65 로 설정한다. 디스플레이의 최대 휘도는 70~250 cd/m² 범위에서 설정하며, 대비비는 0.02 이하를 유지해야 한다. 화면 뒤 배경 휘도의 비율은 영상 최대 휘도의 약 0.15 로 맞춘다.
    - **가정 환경**: 소비자 측면의 품질 평가를 위해 설정되며, 화면에 도달하는 조도는 약 200 lux 수준이다. 디스플레이 휘도는 70~500 cd/m² 범위에서 설정하며, 화면의 비활성 상태 휘도는 최대 휘도의 0.05 이하로 유지한다.
    -
- 응용프로그램 특화 조건 (8 장 기준)

  특정 서비스나 시스템 환경에 맞는 조건을 적용한다. 예를 들어:

    - SDTV 평가: 실험실 환경에서는 시청 거리를 화면 높이의 4H~6H 로 하고, 디스플레이 크기는 최소 20 인치 이상이어야 한다. 가정 환경에서는 4:3 화면비 기준 25~29 인치, 16:9 기준 32~36 인치 크기의 SDTV 화면을 사용하며, 휘도는 200 cd/m²로 유지한다.
    - HDTV 평가: HDTV 조건에서는 Part 1 의 일반 조건을 기본으로 적용하되, 표준화된 화면 크기와 시청 거리(예: 6H)를 고려한다.
    - **멀티미디어 및 모바일 환경**: 시청 거리는 1~8H 범위 내에서 제한되거나 시청자의 선호에 따라 결정할 수 있으며, 배경 조도는 20 lux 이하로 유지한다.

즉, SSCQE 를 포함한 주관적 화질 평가에서 시청 조건은 기본적으로 Part 1 의 실험실·가정 환경 기준을 따르며, 필요할 경우 6 장에서 제시하는 서비스·응용프로그램 특화 조건을 적용하도록 규정되어 있다.

7.5.4 등급 척도

피험자는 핸드셋 슬라이더 메커니즘의 이동 범위가 7.1.4 절에 정의된 연속 품질 척도(Continuous Quality Scale)에 해당함을 사전에 안내 받아야 한다. 이를 통해 평가 점수가 일관되게 수집되도록 한다. 연속 품질 척도는 주관적 화질 평가에서 사용되는 비연속적 등급 척도가 아닌 연속적 값에 기반한 평가 방법을 의미한다. 첨부된 문서에 따르면, 이 척도는 시청자가 영상의 품질을 단일 등급(Excellent, Good 등)으로 구분하지 않고, 연속적인 축 위에서 위치를 선택하는 방식으로 품질을 평가하게 설계되어 있다

- 연속 품질 척도의 형태
    - 전통적인 5 점 척도(Excellent ~ Bad)나 결함 척도 (Imperceptible ~ Very annoying)와 달리, 연속 품질 척도는 선형적이고 연속적인 스케일을 사용한다.
    - 관찰자는 슬라이더(handset slider mechanism)와 같은 장치를 사용하여 자신의 평가를 시간에 따라 연속적으로 기록할 수 있다.
    - 척도의 범위는 일반적으로 0~100 점 또는 유사한 연속 값 구간으로 정의된다.
- 연속 품질 측정 방식
    - 연속 입력 장치: ITU 권고에서는 스프링이 없는 슬라이더 메커니즘, 약 10 cm 이동 범위, 고정식 또는 책상 거치형 장치 사용을 명시한다.
    - 샘플링 속도: 관찰자의 입력은 초당 두 번 이상 기록되어, 시간에 따른 품질 변화를 추적할 수 있다.
    - 지속적 평가: 디지털 영상은 장면에 따라 손상이 매우 짧게 나타나고 사라질 수 있으므로, 단발적인 평가가 아닌 지속적 평가가 필요하다.
- 연속 품질 척도의 의미
    - 척도의 한쪽 끝은 “최고의 품질(Excellent 또는 무결함 상태)”을, 반대쪽 끝은 “최악의 품질(Bad 또는 매우 불쾌한 결함)”을 나타낸다.
    - 사용자는 특정 순간에 체감하는 품질을 이 연속 축 위의 한 지점으로 표현하게 된다.
- 연속 품질 척도 활용 맥락
    - 연속 품질 척도는 특히 SSCQE 에서 핵심적으로 사용된다.
    - 이는 피험자가 소스 참조 영상 없이 단일 자극을 시청하면서 품질을 실시간으로 연속적으로 기록하는 방법이다.

- 결과는 시간 함수 q(t) 형태의 곡선으로 나타나며, 영상 내 품질 변동과 시간적 특성을 분석할 수 있다.

7.5.5 관찰자

관찰자는 최소 15 명 이상의 비전문가(non-expert)로 구성해야 한다. 이는 통계적 신뢰도를 확보하기 위한 최소 요구 조건이다

7.5.6 관찰자 지침

서비스 품질 평가에서 오디오가 포함된 경우, 관찰자는 단순히 영상 품질만이 아니라 오디오와 영상이 결합된 전체적인 품질 경험을 기준으로 평가하도록 지시해야 한다.

7.5.7 데이터 처리 및 결과 제시

모든 테스트 세션에서 수집된 데이터를 기반으로, 프로그램 세그먼트 단위, 품질 매개변수 단위, 혹은 전체 세션 단위로 평균 품질 등급을 계산한다. 이렇게 얻어진 평균 품질 등급은 시간의 함수 q(t) 로 표현할 수 있으며, 이는 그림 7-6 과 같은 형태의 그래프로 제시된다. 이 그림은 특정 코덱(X)을 적용한 상황에서 프로그램 세그먼트 Z 를 시험 조건으로 삼는다

이러한 그래프는 시간에 따른 주관적 화질 변동을 직관적으로 보여주며, 특정 구간에서 발생하는 화질 저하나 회복 현상을 파악하는 데 유용하다

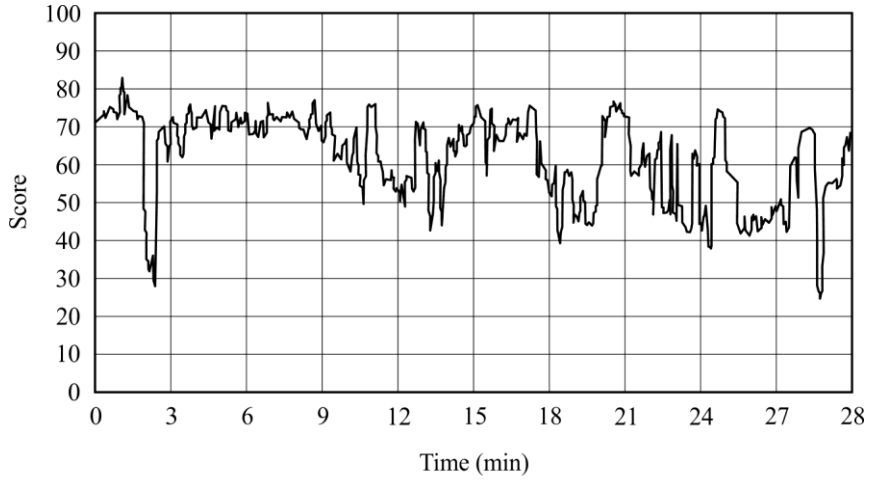


그림 7-6. 시험 조건. 코덱 X / 프로그램 세그먼트: Z

Test condition. Codex X/Programme segment: Z

7.5.8 연속 품질 결과의 보정 및 단일 품질 등급 도출

디지털로 코딩된 비디오에 대한 더 긴 단일 평가 DSCQS 세션에서는 관찰자의 기억 기반 편향(memory-based bias) 이 존재할 수 있음이 확인되었다. 그러나 최근 10 초 길이의 짧은 비디오 발췌본을 대상으로 한 DSCQS 평가에서는 이러한 기억 효과가 유의미하지 않음이 보고되었다.

따라서 현재 연구 중인 SSCQE 프로세스의 두 번째 단계로 고려되는 접근법은 다음과 같다.

- 연속 품질 평가 과정에서 얻은 히스토그램 데이터를 분석한다.
- 여기서 추출된 대표적인 10 초 샘플에 대해 기존 DSCQS 을 적용한다.
- 이를 통해 품질 히스토그램을 보정하고, 연속 평가와 단일 평가 간의 대응 관계를 확립한다.

과거에 사용된 기존 ITU-R 방법론은 비디오 시퀀스 전체에 대해 단일 품질 등급을 도출하는 것이 가능했다. 이에 따라, 최근 연구에서는 코딩된 비디오 시퀀스의 연속적인 품질 평가와 동일한 세그먼트의 단일 품질 등급 간의 관계를 규명하기 위한 실험이 수행되었다.

실험 결과, 시퀀스의 마지막 약 10~15 초 구간에서 눈에 띄는 화질 손상이 발생하는 경우, 인간의 기억 효과가 전체 품질 등급을 왜곡할 수 있음이 확인되었다. 그러나 이러한 기억 효과는 감쇠 지수(exponential decay) 가중 함수를 통해 모델링할 수 있음이

밝혀졌다. 이는 인간이 최근 경험한 품질 열화를 전체적인 평가에 더 강하게 반영한다는 사실을 수학적으로 설명해 준다.

따라서 SSCQE 방법론의 세 번째 단계로 제안되는 것은 다음과 같다.

- 연속 품질 평가 결과를 기억 효과 보정 모델을 적용해 후처리한다.
- 이를 통해 시간에 따른 연속 평가 데이터를 기반으로 ESQS 을 도출한다.

이 접근은 현재 활발히 연구 중이며, 향후 SSCQE 결과를 기존 ITU-R 방법론과 정합성 있게 통합하는 핵심 단계가 될 것으로 기대된다. 그림 7-7 은 프로그램 세그먼트 Z 에서 여러 피험자가 기록한 연속 품질 점수들을 평균화하여 도출한 결과를 보여주는 그래프이며, 나아가 이를 히스토그램으로 변환해 확률적 품질 분포 P(q)를 얻는 과정의 예시이다.

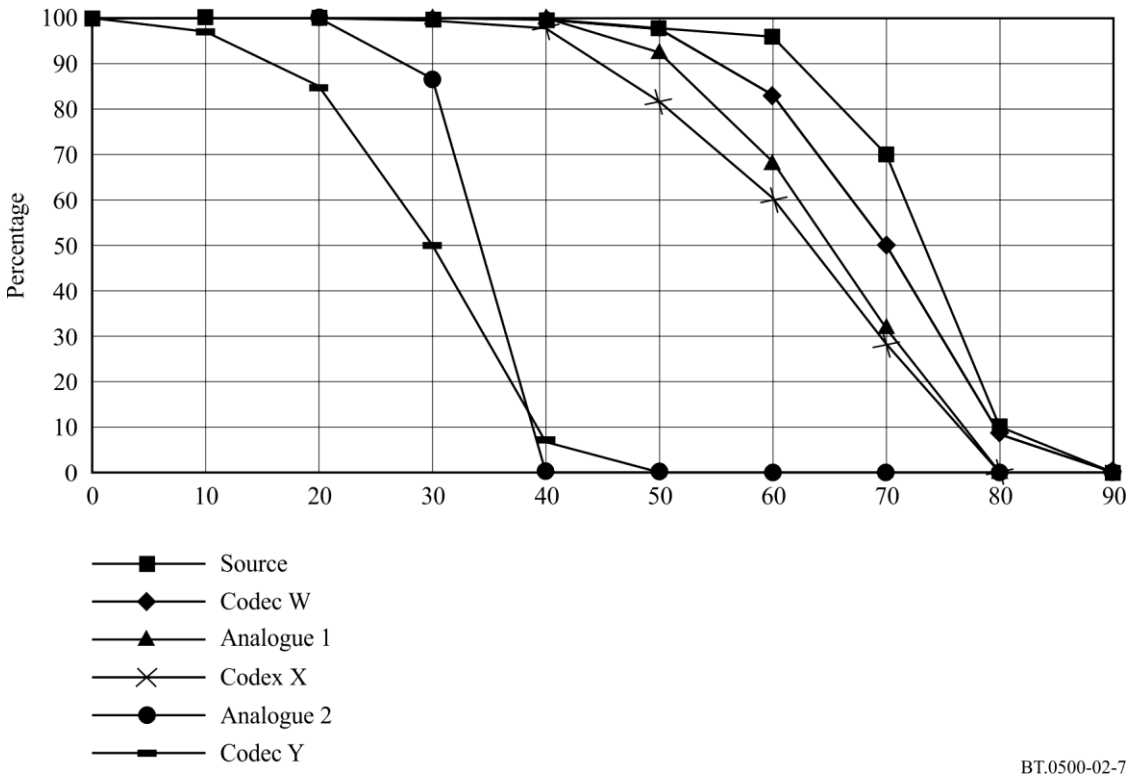


그림 7-7. 프로그램 세그먼트 Z 에 대한 평가 점수의 평균

Mean of scores of voting sequences on programme segment Z

7.6 SDSCE(Single stimulus continuous quality evaluation)

연속 평가(continuous evaluation)라는 개념은 기존의 비디오 품질 평가 방법들이 디지털 압축 방식의 특성을 충분히 반영하지 못한다는 한계에서 출발하였다. 과거 ITU-R 에서 표준화된 기법들은 일반적으로 10 초 정도의 짧은 영상 시퀀스를 대상으로 하였으며, 이는 실제 서비스 환경에서 발생하는 **맥락 의존적 왜곡(artefacts)** 을 반영하기에 충분하지 않았다.

디지털 영상의 손상은 영상의 공간적·시간적 콘텐츠뿐 아니라 압축 방식과 디지털 전송 시스템의 오류 복원 동작에도 크게 의존한다. 따라서 짧은 발췌본으로는 시스템 전체의 성능을 대표적으로 평가하기 어렵다. 이러한 이유로 ITU-R 은 SSCQE 을 도입하여, 더 긴 시퀀스를 대상으로 실제 상황에 가까운 조건에서 품질을 연속적으로 평가할 수 있도록 했다.

그러나 충실도(fidelity)를 평가해야 하는 경우에는 참조(reference) 조건이 필요하다. 이때 SDSCE 는 SSCQE 를 기반으로 발전된 방법으로, 참조 영상과 테스트 영상을 동시에 제시하여 피험자가 **실시간으로 품질 차이를 연속 평가**할 수 있도록 고안되었다. 본 방법은 MPEG 표준에서 매우 낮은 비트레이트 조건에서 오류 강인성을 평가하기 위해 처음 제안되었으나, 시간에 따라 변하는 손상이 충실도에 영향을 주는 다양한 상황에 적용할 수 있다.

7.6.1 테스트 절차

- 피험자는 동시에 두 개의 시퀀스를 시청한다.
    - 전통적인 하나는 참조(reference) 시퀀스,
    - 다른 하나는 테스트(test) 조건 시퀀스이다.
- 시퀀스의 형식이 SIF(Standard Image Format) 또는 그 이하라면 두 시퀀스를 동일한 화면에 나란히 배치할 수 있다.
- 해상도가 더 크다면 정렬된 두 개의 디스플레이를 사용해야 한다 (그림 7-8).

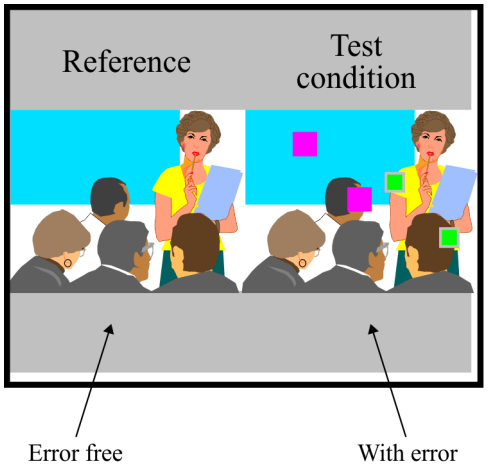


그림 7-8. 디스플레이 포맷 예시

Example of display format

피험자는 두 영상 간의 차이를 관찰하며 핸드셋 슬라이더 장치를 사용하여 충실도를 평가한다.

- 완벽히 동일할 경우 → 척도 최상단(코드 100).
- 전혀 충실도가 없을 경우 → 척도 최하단(코드 0).
- 피험자는 어떤 영상이 참조인지 인지하고 있으며, 시청 내내 지속적으로 평가한다.

7.6.2 테스트 단계 (Phases)

- 1. 훈련 단계
  - 피험자가 과제를 잘못 이해하지 않도록 충분한 설명을 제공한다.
  - 서면 지침에는 평가 대상, 평가 항목(품질 차이), 그리고 평가 방법이 포함되어야 한다.
  - 모든 질문은 편향 없이 명확히 답변해야 한다.
- 2. 시범 세션
  - 피험자들이 절차와 손상의 유형에 익숙해질 수 있도록 시범을 진행한다.
- 3. 모의 테스트(Mock test)
  - 실제 테스트에 사용되지 않는 시퀀스를 활용하여 여러 조건을 제시한다.
  - 참조와 동일한 조건에서는 평가가 100 에 가깝게 나타나는지 확인한다.
  - 피험자가 부적절한 반응을 보일 경우 설명 및 모의 테스트를 반복한다.

7.6.3 테스트 프로토콜 특징

- **비디오 세그먼트 (VS)**: 하나의 비디오 시퀀스.
- **테스트 조건 (TC)**: 특정 비디오 처리 과정, 전송 조건, 또는 두 가지의 조합. 각 VS 는 최소 하나의 TC 에 따라 처리된다.
- **세션 (S)**: 여러 VS/TC 쌍을 의사 무작위 순서로 제시한 연속 시청 구간.
- **테스트 제시 (TP)**: 모든 VS/TC 조합을 포함하는 일련의 세션. 동일한 관찰자가 아니더라도 모든 조합은 균등하게 평가되어야 한다.
- **투표 기간**: 관찰자는 세션 전체 동안 연속적으로 평가를 기록한다.
- **투표 세그먼트 (SOV)**: 10 초 분량의 연속 평가 단위. 각 SOV 는 20 개의 투표 샘플(0.5 초 간격 기록)에 해당한다.

7.6.4 데이터 처리

SDSCE 절차로 수행된 테스트에서 수집된 데이터는 모든 세션(S)에 걸쳐 관찰자들의 투표 결과로 구성된다. 데이터의 유효성 검증은 각 VS/TC 조합이 균등하게 평가되었는지를 확인하는 것으로 시작한다. 이후 다단계 분석이 필요하다:

- **1. 개별 투표 단위 분석**
    - 모든 관찰자의 연속 평가를 누적하여 각 시점별 평균과 표준편차를 산출한다.
- **2. 투표 세그먼트(SOV) 단위 분석**
    - 그림 7-9 에 설명된 절차에 따라, 20 개의 연속 투표(10 초 구간)를 묶어 각 SOV 의 평균과 표준편차를 계산한다.
    - 이 결과는 시간에 따른 품질 변화를 보여주는 다이어그램으로 시각화되며, 그림 7-10 에서 예시된다.
- **3. 분포 및 출현 빈도 분석**
    - 각 SOV 평균값을 기반으로 통계적 분포와 출현 빈도를 계산한다.
    - 이전 VS × TC 조건의 영향(특히 **최신 효과, recency effect**)을 줄이기 위해 각 조합의 첫 10 개 SOV 는 제외한다.
- **누적 성가심(annoyance) 특성 도출**
    - 출현 빈도를 누적하여 전체 성가심 특성을 계산한다.
    - 이때 신뢰 구간을 반드시 고려해야 하며, 결과는 그림 7-11 과 같은 누적 분포 함수 형태로 표현된다.

○ 이렇게 얻어진 누적 통계 분포 함수는 각 투표 세그먼트의 평균과 그 발생 빈도 간의 관계를 나타내며, 서비스 전반의 품질 안정성과 열화 특성을 파악하는 데 활용된다.

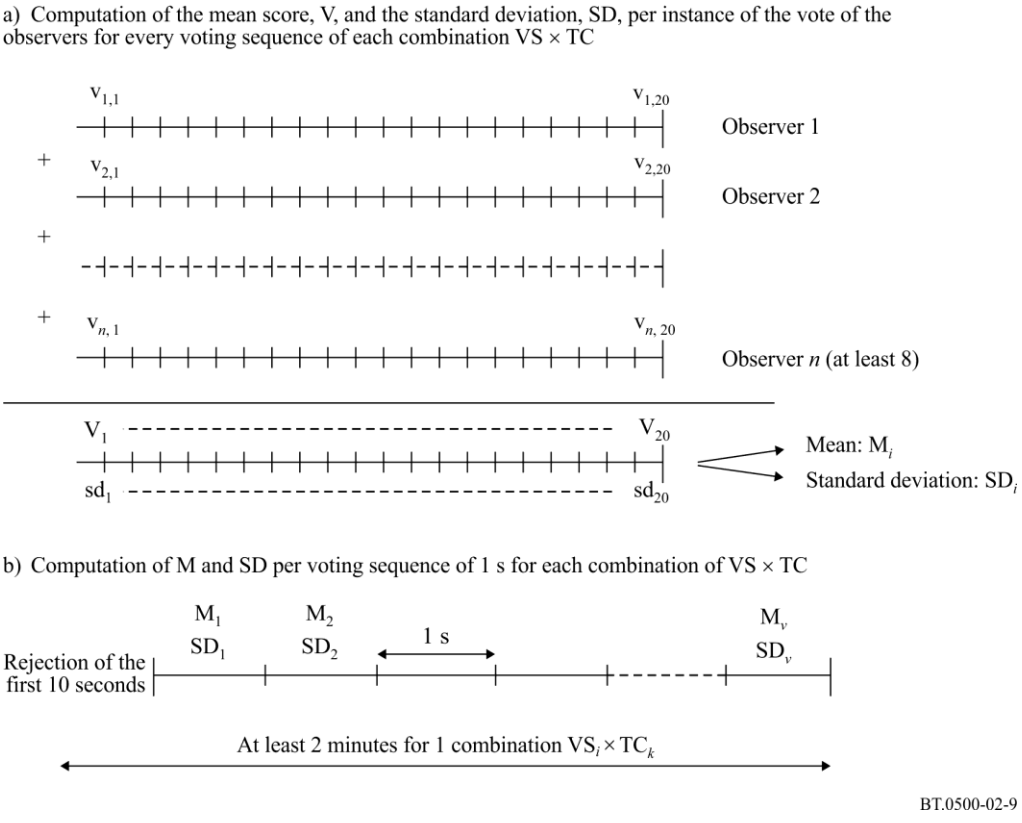


그림 7-9. 데이터 프로세싱

Data processing

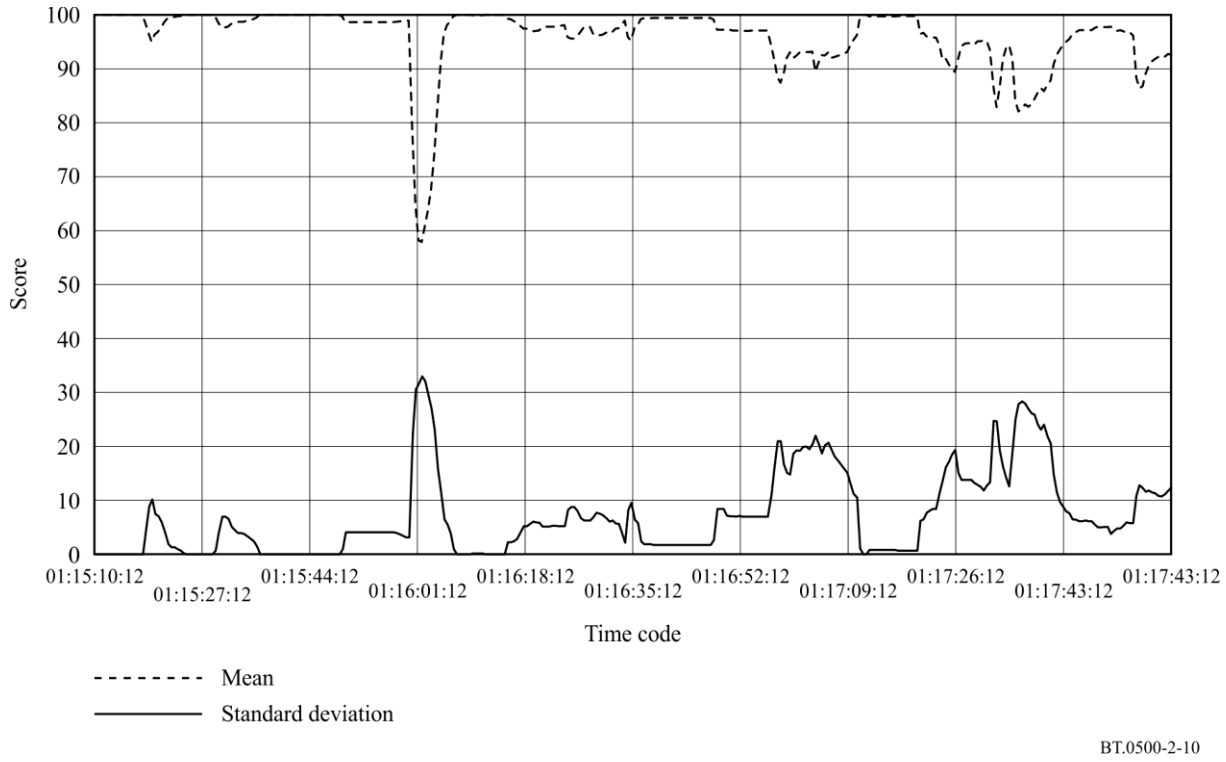


그림 7-10. 투표 세그먼트(SOV)별 통계 매개변수의 시간 다이어그램

Raw temporal diagram

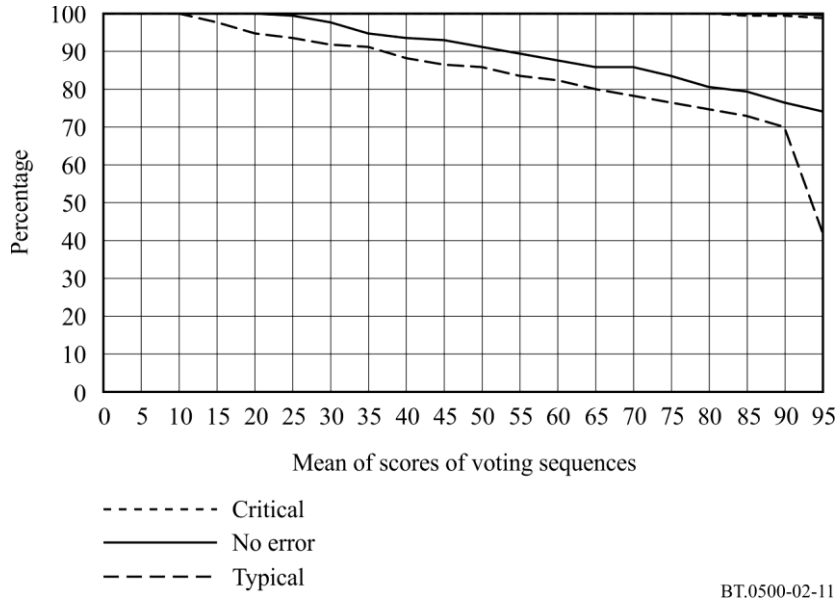


그림 7-11. 신뢰 구간을 포함한 누적 성가심 특성

Global annoyance characteristics calculated from the statistical distributions and including confidence interval

7.6.5 피험자의 신뢰도 평가

피험자의 신뢰도는 다음 방식으로 검증한다.

- 참조/참조 쌍이 제시될 경우 평가값이 100 에 근접하는지 확인한다. 이는 피험자가 과제를 이해하고 무작위로 반응하지 않음을 보여준다.
- SSCQE 에서 사용되는 관찰자 신뢰도 검증 절차 6.7.2.3.2 를 적용할 수 있다.

신뢰도 저하 요인과 대응 방안은 다음과 같다.

- **체계적 이동(Systematic shift)**
  - 피험자가 지나치게 낙관적이거나 비관적으로 반응하거나, 혹은 투표 절차 자체(예: 척도의 의미)를 잘못 이해할 경우 발생한다.
  - 이 경우, 전체적인 평균 범위를 크게 벗어나지는 않더라도, 평가 곡선이 평균보다 지속적으로 더 높거나 낮게 이동된 상태로 나타난다.
  - 이는 평가자의 성향이나 착오로 인해 발생하는 지속적 편향(bias)에 해당한다.
- **국소적 역전(Local inversion)**
  - 다른 테스트 절차에서도 종종 나타나듯, 피험자가 순간적으로 집중을 덜 하거나 영상 품질 추적을 소홀히 할 때 발생한다.
  - 전반적인 평가 곡선은 평균 범위 내에 있더라도, **특정 구간에서 실제 품질과 반대로 평가하는 국소적 왜곡**이 나타날 수 있다.
  - 즉, 짧은 구간에서 “품질이 좋아졌는데 나빠졌다” 또는 “나빠졌는데 좋아졌다”와 같이 잘못된 반응이 기록될 수 있다.
- **예방 및 대응**
  - 이러한 **비정형적 행동과 역전 현상**은 훈련 과정에서 어느 정도 예방할 수 있다.
  - 그러나 훈련만으로는 완벽히 배제하기 어렵기 때문에, 문서는 **불일치한 관찰자를 탐지하고 필요한 경우 제외할 수 있는 도구**의 사용을 권장한다.
  - 이를 위해 ITU-R 권고에서는 **2 단계 필터링 프로세스**를 제안하고 있으며,
    - **1 단계**: 평균적 경향과 크게 다른 체계적 이동을 보이는 관찰자 제거
    - **2 단계**: 체계적 이동은 없지만, **국소적 불일치**가 많은 관찰자 제거의 방식으로 평가 신뢰도를 관리한다

7.7 SAMVIQ(Subjective Assessment Method for Video Quality in multimedia applications)

7.7.1 서론

SAMVIQ 는 멀티미디어 환경에서 비디오 시퀀스의 **내재적 품질**을 측정하기 위해 고안된 주관적 평가 기법이다. 이 방법은 **연속 품질 척도(0~100)** 를 사용하며, 척도에는 5 개의 품질 항목(매우 좋음, 좋음, 보통, 나쁨, 매우 나쁨)이 선형적으로 주석 처리되어 있어 관찰자가 직관적으로 판단할 수 있도록 설계되어 있다.

SAMVIQ 의 특징은 다음과 같다.

- 관찰자는 하나의 장면(Scene)에 대해 여러 버전의 시퀀스를 무작위로 접근할 수 있다.
- 각 버전은 명시적 참조(explicit reference), 숨겨진 참조(hidden reference), 처리된 시퀀스(processed sequence)를 포함한다.
- 관찰자는 각 버전을 자유롭게 **재생, 정지, 반복 시청, 점수 수정**할 수 있으며, 평가를 원하는 방식으로 조정할 수 있다.
- 이는 단일 자극 방법(SS)의 자유도와, 참조 기반 방법(DSCQS)의 정밀성을 모두 결합한 **하이브리드형 평가 방법**이라 할 수 있다

이 방법을 더 잘 이해하기 위해 다음과 같은 특정 용어를 정의한다

- Scene: 시청각 콘텐츠
- Sequence: 처리가 결합되었거나 처리되지 않은 장면
- Algorithm: 하나 또는 여러 개의 이미지 처리 기술

7.7.2 명시적 참조, 숨겨진 참조 및 알고리즘

- SAMVIQ 는 결과의 안정성을 확보하기 위해 품질 앵커(anchor) 를 포함한다.
- 두 개의 고품질 앵커가 권장되며, 명시적 참조는 점수의 표준 편차를 줄이고 신뢰성을 극대화한다.
- 그러나 약 30%의 관찰자는 명시적 참조에 무조건 최고 점수(100)를 주는 경향이 있어, 숨겨진 참조를 도입하여 내재적 품질을 평가할 수 있도록 설계되었다
- SAMVIQ 는 코덱 유형, 비트레이트, 이미지 형식, 시간적 업데이트, 줌 효과 등의 다양한 처리 특성을 결합할 수 있으며, 이러한 특성 또는 조합을 알고리즘(Algorithm) 이라 부른다.

7.7.3 테스트 조건

- SAMVIQ에서는 다른 방법론(예: 단일 자극 방법)에서 암묵적으로 사용되는 규칙에 따라 동질적인 콘텐츠를 선택한다.
- 이는 장면(Scene) 내에서 중요도(criticality) 변화를 최소화하여, 안정적이고 신뢰할 수 있는 품질 점수를 얻을 수 있도록 한다.
- 시퀀스 길이는 10 초 또는 15 초로 제한되며, 이는 충분히 신뢰할 만한 결과를 제공한다.
- 적절한 디스플레이 성능을 유지하기 위해 전용 디코더-플레이어 또는 그 출력 사본을 사용해야 한다.

7.7.4 테스트 구성

SAMVIQ 의 테스트 절차는 그림 7-12 의 구조에 따라 이루어진다.

- (a) 테스트는 장면(Scene) 단위로 수행된다.
- (b) 관찰자는 현재 장면의 시퀀스를 원하는 순서로 재생하고 점수를 부여할 수 있으며, 반복 평가도 가능하다.
- (c) 장면 간 시퀀스 접근은 무작위화되며, 동일한 순서로 반복 투표하는 것을 방지한다. 단, 알고리즘 순서는 동일하게 유지된다.
- (d) 첫 시청 시에는 반드시 전체 시퀀스를 시청한 후 평가해야 한다.
- (e) 다음 장면으로 이동하기 위해서는 현재 장면의 모든 시퀀스에 점수를 부여해야 한다.
- (f) 전체 테스트 종료 조건은 모든 장면의 모든 시퀀스를 평가 완료하는 것이다..

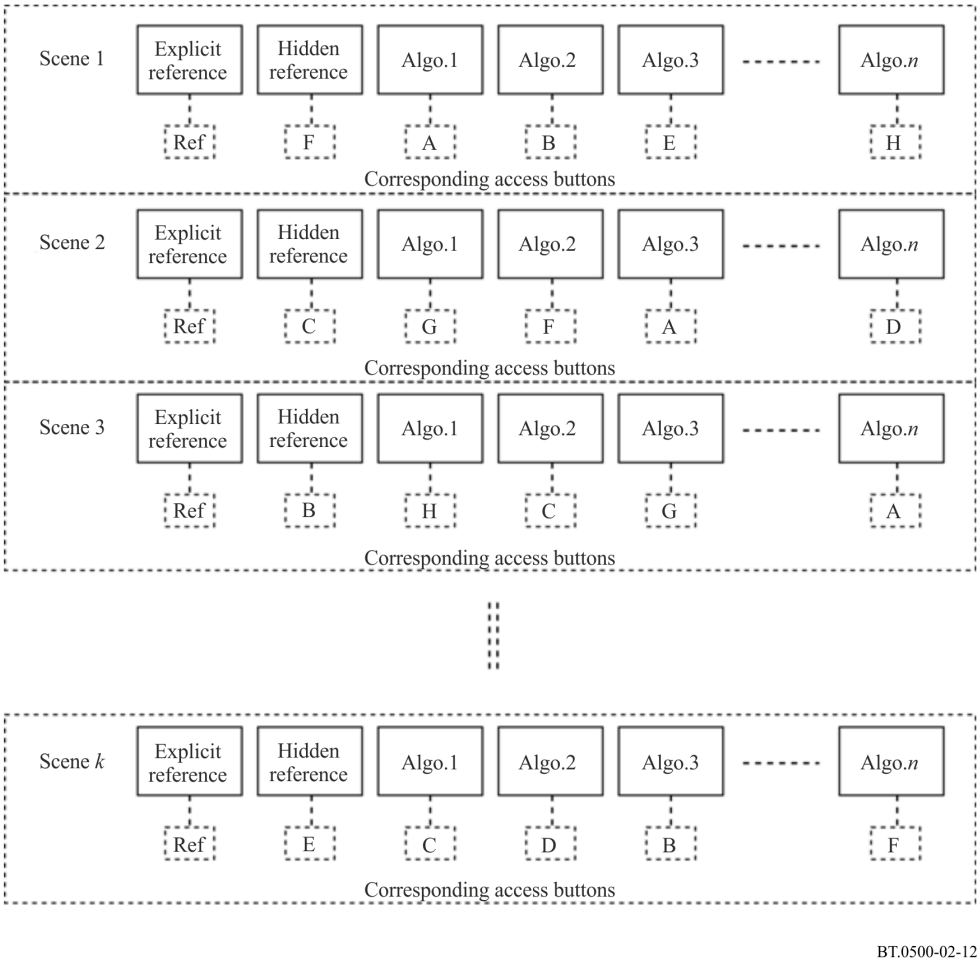


그림 7-12. SAMVIQ 방법의 테스트 구성 예시

Test organization example for the SAMVIQ method

소프트웨어 구현

- 그림 7-12 에 따라 인터페이스가 구성된다.
- 관찰자는 “재생, 정지, 다음 장면, 이전 장면” 버튼을 사용하여 시퀀스를 제어한다.
- 점수는 접근 버튼 아래에 표시되며, 모든 시퀀스를 평가한 뒤에도 점수 수정이 가능하다.
- 큰 차이는 첫 시청에서 이미 확인되므로 전체 재시청은 필요하지 않다.

7.7.5 데이터 제시 및 분석

7.7.5.1 요약 정보

테스트 환경에 대한 정보는 결과 비교와 재현성을 위해 보고해야 한다. (표 7-3 참조) 필수 항목: 디스플레이 기술, 최대/흑색 휘도, 배경 휘도, 조명(lux), 시청 거리, 화면 크기, 입력·출력 형식, 화이트 포인트 좌표, 관찰자 수 등.

표 7-3. 테스트 요약 정보
Test summary information

Name of the method	
Display technology	
Reference name of the display	
Peak luminance level (cd/m²)	
Black luminance level (cd/m²)	
Black level setup: PLUGE (black to supra black level distance perceived threshold = 8). Otherwise indicates the threshold value	
Background luminance level (cd/m²)	
Illumination (lux)	
Viewing distance: – Not constrained: front of display – Constrained: nH	
Display size (diagonal in inches)	
Width/height display ratio	
Display format (number of columns and lines)	
Image input format (number of columnsand lines)	
Image output format[6] (number of columns and lines)	
White colour temperature: D_{65} otherwise White colour coordinates (x, y)	

[6] 이 정보는 입력 이미지가 디스플레이에서 처리될 때(예: 리스케일링) 필요합니다.

Number of effective observers	

평판 디스플레이라면 감마 충실도와 색상 원색 정보도 포함해야 한다. 비디오 시퀀스의 시공간적 특성 보고 역시 필요하다.

7.7.5.2 분석 방법

분석은 6.7.2 절에 제시된 방법론을 따른다.

6.7 절은 MOS 와 신뢰 구간 계산, 관찰자 신뢰도 검증, 데이터 보정, 함수 근사화 등 데이터 분석의 표준 프로토콜을 제공한다. 이를 통해 실험 결과의 신뢰성과 재현성을 확보하며, 서로 다른 연구 간 비교를 가능하게 만든다.

7.7.5.3 관찰자 선별

SAMVIQ 의 관찰자 선별은 6.7.2.3.3 절에 규정된 절차를 따른다. 이 과정은 신뢰도 높은 관찰자만 데이터를 반영하도록 보장한다.

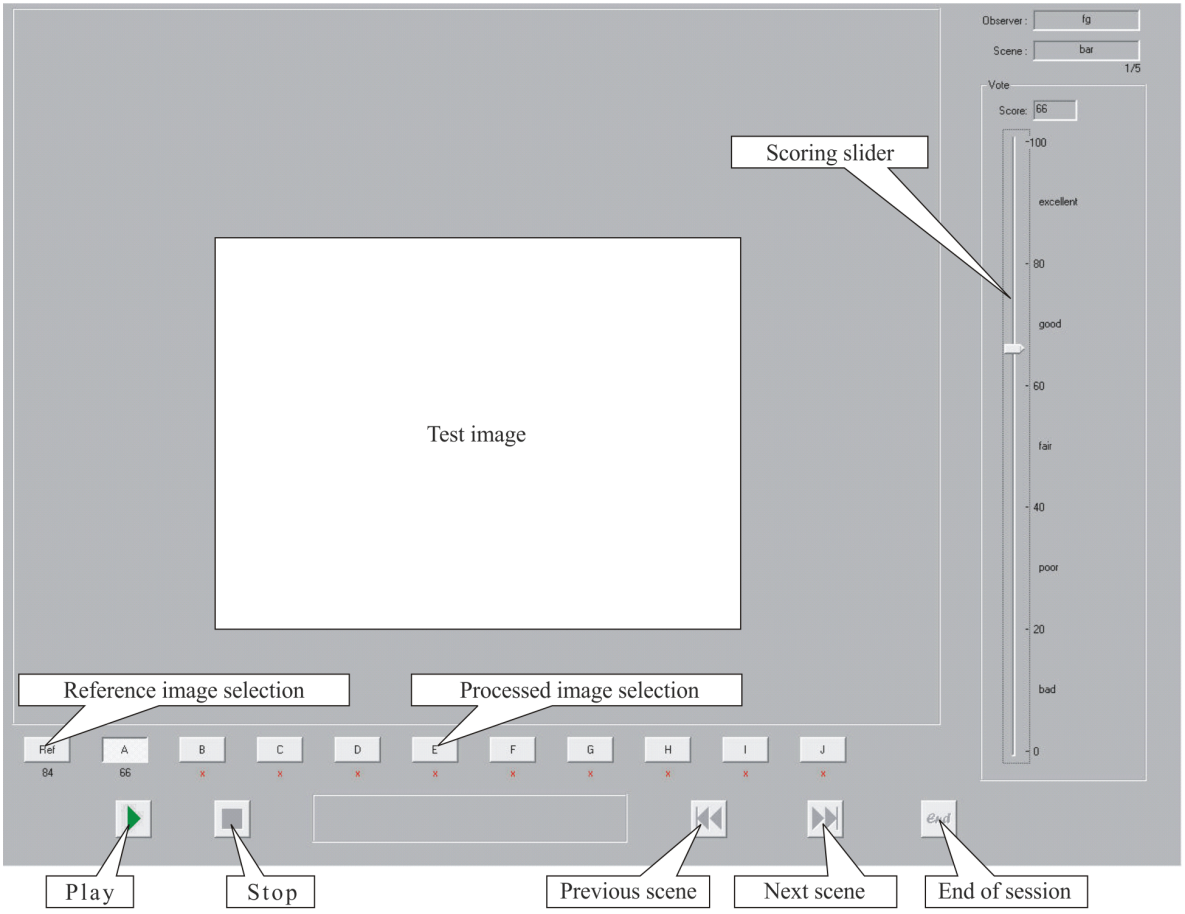
6.7.2.3.3 절은 상관관계 기반 관찰자 선별 절차를 규정한다. 피어슨 및 스피어만 상관계수를 이용하여, 개별 평가가 집단 평균과 충분히 일치하지 않을 경우 해당 관찰자를 제거한다. 이 과정은 주관적 평가 데이터의 신뢰성, 일관성, 재현성을 확보하는 핵심 절차이다.

7.7.6 SAMVIQ 인터페이스 예시 (참고용)

이 그림은 관찰자가 **각 시퀀스 버전을 선택하고 평가하는 그래픽 사용자 인터페이스**의 구조를 보여준다. 인터페이스에는 다음과 같은 주요 요소가 포함된다.

- **접근 버튼(Access buttons)**: 각 시퀀스를 재생할 수 있는 버튼으로, 관찰자는 이 버튼을 눌러 해당 버전의 시퀀스를 자유롭게 재생·정지할 수 있다.
- **재생·정지·장면 이동 버튼**: "재생(Play)", "정지(Stop)", "다음 장면(Next Scene)", "이전 장면(Previous Scene)" 버튼을 통해 장면 전환 및 제어가 가능하다

- **평가 표시(Score display)**: 관찰자가 점수를 입력하면, 해당 점수는 관련된 버튼 아래 표시됩니다. 모든 시퀀스 버전에 점수가 부여된 이후에도 관찰자는 점수를 비교하고 필요시 수정할 수 있다.
- **연속 품질 척도(Continuous scale)**: 각 시퀀스는 개별적으로 표시되며, 0~100 범위의 연속 품질 척도를 통해 평가된다.



BT.0500-02-12a

7.8 EVP(Expert viewing protocol)

비디오 처리 분야의 전문가 중에서 선택된 소수의 시청자 참여를 통해 동영상의 주관적 화질을 평가하는 방법을 설명한다.

7.8.1 실험실 설정

7.8.1.1 디스플레이 선택 및 설정

사용되는 디스플레이는 전문적인 애플리케이션(예: 방송 스튜디오 또는 중계차)에 일반적으로 사용되는 평판 디스플레이여야 한다. 디스플레이 대각선 크기는 최소 22 인치에서 40 인치(권장)까지 다양할 수 있으며, HDTV 이상의 해상도를 가진 이미지 시스템이 평가될 때는 50 인치 이상으로 확장될 수 있다.

디스플레이의 활성 시청 영역 중 축소된 부분을 사용할 수 있다. 이 경우 디스플레이의 활성 부분 주변 영역은 중간 회색으로 설정되어야 한다. 이러한 사용 조건에서는 디스플레이가 원래 해상도와 다른 해상도로 설정되어서는 안 된다.

디스플레이는 전문적인 측광기를 사용하여 휘도와 색상에 대해 적절한 설정 및 보정을 거쳐야 한다. 디스플레이의 보정은 수행 중인 테스트에 대해 관련 권고안에서 지정된 매개변수를 준수해야 한다.

7.8.1.2 시청 거리

전문가가 시청해야 하는 거리는 디스플레이의 해상도와 활성 부분의 높이에 따라 선택되어야 하며, 6.1.3.2 절에 기술된 설계 시청 거리 또는 중요한 시청 조건의 요구사항에 따라 더 짧은 시청 거리를 따라야 한다.

7.8.1.3 시청 조건

EVP 실험은 반드시 테스트 실험실에서 진행될 필요는 없지만, 테스트 위치가 청각적 또는 시각적 방해로부터 보호되어야 한다. (예로, 조용한 사무실이나 회의실이 사용될 수 있다).

화면에 직접 또는 반사되는 조명은 제거되어야 한다. 다른 주변 조명은 낮게 유지되어야 하며, (조명을 사용하는 경우) 채점 시트를 채울 수 있는 최소 수준으로 유지되어야 한다.

디스플레이 앞에 앉는 전문가의 수는 화면 크기에 따라 달라질 수 있으며, 모든 시청자에게 동일한 이미지 렌더링과 자극 제시를 보장해야 한다.

7.8.2 시청자

EVP 에 참여하는 시청자는 연구 분야의 전문가여야 한다. 시청자는 자격 있는 인원으로 선택되므로, 시력이나 색맹에 대해 고려하여 시청자를 선별할 필요는 없다. 서로 다른 최소 시청자 수는 9 명이어야 한다.

최소 시청자 수에 도달하기 위해, 동일한 실험을 여러 장소에서 테스트를 반복하여 수행할 수 있다. 다른 장소에서 EVP 세션에 참여하는 시청자들의 점수는 통계적으로 함께 처리될 수 있다.

7.8.3 기본 테스트 셀

전문가에게 제시될 자료는 평가할 각 코딩 조건 쌍에 대해 기본 테스트 셀(BTC: basic test cell)을 생성하여 구성되어야 한다 (그림 7-13 참조).

BTC 에서 고려할 소스 참조 시퀀스(SRC)와 처리된 비디오 시퀀스(PVS) 클립은 전문가가 테스트 중인 압축 알고리즘에 의해 제공되는 시각적 품질의 개선 사항을 식별할 수 있도록 항상 동일한 비디오 시퀀스와 관련되어야 한다.

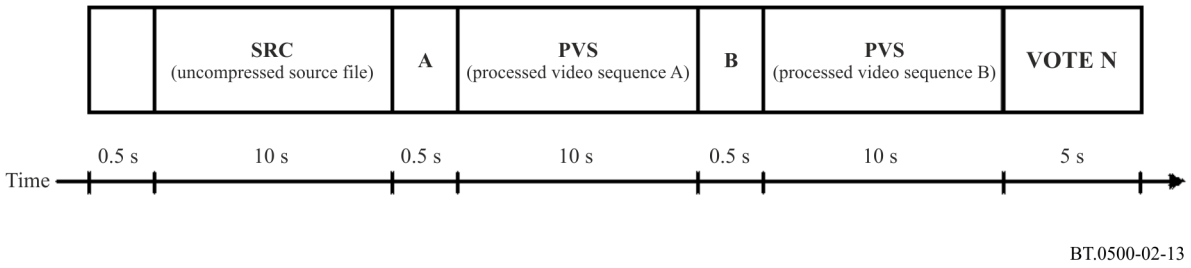


그림 7-13. EVP에서의 BTC 타이밍

TIMINGS OF A BASIC TEST CELL FOR THE EXPERT VIEWING PROTOCOL

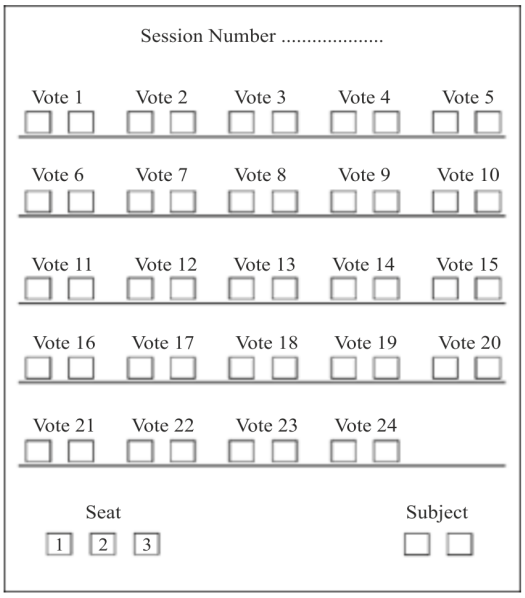
BTC 는 다음과 같이 구성되어야 한다:

- **0.5초** 동안 화면을 중간 회색으로 설정 (휘도 척도의 평균값);
- **10초** 동안 참조용 비압축 비디오 클립 제시;
- **0.5초** 동안 중간 회색 배경에 메시지 "A"(첫 번째 평가 비디오) 표시;
- **10초** 동안 손상된 버전의 비디오 클립 제시;
- **0.5초** 동안 중간 회색 배경에 메시지 "B"(두 번째 평가 비디오) 표시;
- **10초** 동안 손상된 버전의 비디오 클립 제시;
- **5초** 동안 시청자에게 의견을 표현하도록 요청하는 메시지 표시.

'투표' 메시지 다음에는 채점 시트와 동기화하는 데 도움이 되는 번호가 와야 한다.

7.8.4 채점 시트 및 평가 척도

그림 7-13 에 표시된 바와 같이, 비디오 클립의 제시는 손상되지 않은 참조(SRC)가 먼저 표시된 다음 두 개의 손상된 비디오 시퀀스(PVS)가 표시되는 방식으로 배열되어야 한다. PVS 의 제시 순서는 각 BTC 마다 무작위로 변경되어야 하며 시청자는 제시 순서를 알아서는 안 된다.



BT.0500-02-14

그림 7-14. 24-BTC 전문가 평가 세션의 평가표 예시

EXAMPLE OF SCORING SHEET FOR A 24-BTC EXPERT VIEWING SESSION

10(감지할 수 없는 손상)에서 0(매우 성가신 손상)까지의 11 등급 숫자 척도가 사용된다.

표 7-4 는 11 등급 숫자 척도의 의미에 대한 지침을 제공한다.

표 7-4. 11 단계 수치 척도의 의미

Meaning of the 11 grades numerical scale

Score	**Impairment item**	
10	Imperceptible	
9	Slightly perceptible	somewhere
8		everywhere
7	Perceptible	somewhere
6		everywhere
5	Clearly perceptible	somewhere
4		everywhere
3	Annoying	somewhere
2		everywhere
1	Severely annoying	somewhere
0		everywhere

시청자는 각 BTC 에 대해 "A"와 "B"로 표시된 두 개의 상자로 구성된 설문지를 작성하도록 요청 받으며, 두 상자 각각에 11 등급 숫자 척도에서 선택한 점수를 기입해야 한다.

그림 7-14 는 24 개의 BTC 로 구성된 세션의 채점 시트 예시를 제공한다.

각 BTC 에 대해, 시청자는 문자 A 로 식별된 상자(첫 번째로 제시된 비디오 클립 평가용)와 문자 B 로 식별된 상자(두 번째로 제시된 비디오 클립 평가용)를 모두 채운다. 원본의 손상되지 않은 비디오 클립을 제시함으로써 전문가들은 모든 손상을 더 쉽게 평가할 수 있다.

11 등급 숫자 척도의 의미는 아래에 기술된 "훈련 세션" 동안 신중하게 설명되어야 한다.

7.8.5 테스트 설계 및 세션 생성

BTC 의 제시 순서는 테스트 설계자에 의해 동일한 비디오 클립이나 동일한 손상된 클립이 연속으로 두 번 제시되지 않는 방식으로 무작위 순서로 설정되어야 한다.

모든 시청 세션은 각 테스트 세션에 포함된 것들 중 "최상", "최악" 및 두 개의 "중간 품질" BTC 를 포함하는 "안정화 단계"로 시작해야 한다. 이를 통해 시청자들은 테스트 세션 시작부터 품질 범위에 대한 즉각적인 인상을 가질 수 있다.

만약 시청 세션이 20 분보다 길 경우, 테스트 설계자는 이를 각각 20 분을 초과하지 않는 두 개 (이상)의 별도 시청 세션으로 나누어야 한다. 이 경우, 각 시청 세션 전에 "안정화 단계"가 제공되어야 한다.

7.8.6 훈련

이 절차가 전문가의 참여를 예상하고 있더라도, 각 실험 전에 짧은 (5-6 BTC) 훈련 시청 세션을 구성하는 것이 바람직하다.

훈련 세션에 사용되는 비디오 자료는 실제 세션 동안 사용될 것과 동일할 수 있지만, 제시 순서는 달라야 한다.

시청자들은 11 단계 척도의 사용법을 익히기 위해, 화면에 “A”와 “B”라는 메시지가 표시된 직후에 제시되는 비디오 클립을 주의 깊게 보고, 처음에 제시된 비디오 클립(SRC)과 차이가 있는지 확인하도록 해야 한다.

7.8.7 데이터 수집 및 처리

점수는 각 세션이 끝날 때 수집되어야 하며, 평균(MEAN) 값을 계산하기 위해 전자 스프레드시트에 기록되어야 한다.

선형 피어슨 상관관계(linear Pearson’s correlation)를 사용하여 시청자에 대한 '사후 선별'을 수행하는 것이 바람직하다. '상관관계' 함수는 평균 의견 점수(MOS)와 관련하여 각 피험자의 모든 점수를 고려하여 적용되어야 한다. 각 시청자를 '수용 가능' 또는 '거부'로 정의하기 위해 임계값을 설정할 수 있다 (권고안 ITU-T P.910 은 0.75 의 폐기 임계값 미만의 피험자를 거부할 것을 제안한다).

7.8.8 EVP 결과의 사용 조건

EVP 는 공식적인 주관적 평가 실험을 수행하기 위한 시간과 자원이 허용되지 않을 때 사용될 수 있다.

EVP 는 공식적인 주관적 평가보다 시간이 덜 걸리며, 실행되는 주변 환경이 시각적 및 청각적 외부 방해로부터 보호된다는 가정 하에 '비공식적인' 환경에서 실행될 수 있다.

유일한 필수 조건은 위 단락에서 설명된 바와 같이 주변 조명 및 시청 조건(디스플레이, 관찰 각도 및 시청 거리)과 관련이 있다.

7.8.9 EVP 결과 사용의 한계

EVP 가 단 9 명의 시청자만으로도 수용 가능한 결과를 제공할 수 있음을 보여준다 하더라도, EVP 실험으로 제공된 MOS 는 공식적인 주관적 평가 실험으로 얻을 수 있는 결과를 대체하는 것으로 간주될 수 없다.

EVP 를 사용하여 얻은 MOS 데이터는 손상 수준에 대한 예비적인 지표를 얻는 데 사용될 수 있다.

EVP 를 사용하여 얻은 MOS 데이터는 평가 중인 비디오 처리 방식의 예비 순위를 매기는 데 사용될 수 있다.

편리하거나 필요하다고 판단되는 경우, 시청 조건, 시청 거리 및 테스트 설계가 동일하다고 가정하고 여러 장소에서 EVP 실험을 병렬로 실행할 수 있다.

동일한 EVP 실험에 참여하는 전문가 시청자의 수가 (실험을 다른 장소에서 실행하는 경우를 포함하여) 15 명 이상인 경우, 원시 주관적 데이터를 처리하여 MOS, 표준 편차 및 신뢰 구간 데이터를 얻을 수 있으며, 이는 테스트 중인 사례에 대한 더 정확한 순위를 매기는 데 도움이 될 수 있다. 이 경우, 더 정확한 추론 통계 분석(예: T-검정)이 수행될 수 있다.

8 응용분야에 특화된 주관적 평가 방법론

8.1 SDTV

표준 화질(SDTV) 텔레비전 시스템의 주관적 평가

8.1.1 서론

기존 텔레비전 시스템의 품질 수준에 있거나 그에 근접한 품질 수준을 제공하는 디지털 시스템의 주관적 평가에 대한 권고안의 일반적인 방법 적용에 관한 세부 사항을 제시한 것이다. 본 문서에 제시된 절차적 세부 사항은 관련 배경 정보와 함께, 기여 및 배포 응용프로그램에서 ITU-R BT.601 에 따라 제작된 자료를 전달하는 데 사용되는 코덱(또는 시스템) 테스트 뿐만 아니라 방송 애플리케이션에서 사용되는 것에도 관련되는 것이다.

배포 애플리케이션의 경우, 품질 사양은 관찰자의 주관적 판단으로 표현될 수 있는 것이다. 따라서 이러한 코덱은 이론적으로 이러한 사양에 대해 주관적으로 평가될 수 있는 것이다. 그러나 기여 응용프로그램으로 설계된 코덱의 품질은 그 출력이 즉각적인 시청이 아닌 스튜디오 후처리, 저장 및/또는 추가 전송을 위한 코딩을 목적으로 하기 때문에 이론적으로 주관적 성능 매개변수로 지정할 수 없는 것이다. 다양한 후처리 작업에 대해 이 성능을 정의하기 어렵기 때문에, 선호되는 접근 방식은 실제적인 기여 응용프로그램을 대표한다고 간주되는 후처리 기능을 포함한 장비 체인의 성능을 지정하는 것인 것이다. 이 장비 체인은 일반적으로 코덱, 그 다음에 스튜디오 후처리 기능(또는 기본 기여 품질 평가의 경우 다른 코덱), 그리고 신호가 시청자에게 도달하기 전에 또 다른 코덱으로 구성될 수 있는 것이다. 기여 응용프로그램 코덱 사양에 대한 이 전략을 채택한다는 것은 본 권고안에 제시된 측정 절차가 그것들을 평가하는 데에도 사용될 수 있음을 의미하는 것이다.

주관적 평가 분야에서는 많은 경험이 축적되어 있으며, 이에 따라 테스트 조건과 방법론을 추천할 수 있는 것이다. 그러나 품질 또는 손상 목표를 지정할 때 기존 방법들은 절대적인 주관적 등급을 제공하는 것이 아니라 참조 및/또는 앵커 조건의 선택에 의해 어느 정도 영향을 받는 결과를 제공한다는 점을 유념해야 하는 것이다. 동일한 방법론이 고정 및 가변 단어 길이 코덱, 그리고 인트라필드 및 인터프레임 코덱 모두에 채택될 수 있으나, 테스트 이미지 시퀀스의 선택은 영향을 받을 수 있는 것이다.

고품질 코덱의 순위를 평가하는 가장 신뢰할 수 있는 방법은 동일한 조건 하에서 모든 후보 시스템을 동시에 평가하는 것인 것이다. 미세한 품질 차이가 관련된

독립적으로 수행된 테스트는 우수성에 대한 논란의 여지가 없는 증거가 아니라 지침으로 사용되어야 하는 것이다.

유용한 주관적 척도는 코더와 디코더 사이의 전송 링크에서 발생하는 비트 오류율의 함수로 결정되는 손상일 수 있는 것이다. 현재로서는 오류 클러스터링이나 버스트를 설명하는 모델의 매개변수를 추천하기에는 실제 전송 오류 통계에 대한 실험적 지식이 부족한 상황이다. 이 정보를 사용할 수 있게 될 때까지 푸아송(Poisson) 분포 오류를 사용하는 것이 가능한 대안인 것이다.

8.1.2 시청조건

디지털 시스템의 주관적 평가를 위한 특정 시청 조건은 다음 단락에 제시한다.

8.1.3 실험실 환경

실험실 환경은 시스템을 점검하기 위한 중요한 조건을 제공하기 위한 것이다. 실험실 환경에서의 주관적 평가를 위한 특정 시청 조건은 표 8-1 에서 제시한다.

표 8-1. 실험실 환경에서 디지털 시스템의 주관적 평가를 위한 특정 시청 조건
Specific viewing conditions for subjective assessments of digital systems in laboratory environment

Condition	Item	Values
a	시청 거리 대 이미지 높이의 비율	4 *H* and 6 *H*[7]
b	최대 휘도	70 cd/m^2
c	사양을 충족하는 배경 부분에 의해 형성되는 시야각	≥43° H × 57° W
d	디스플레이	High quality screen. Size ≥ 20" (50 cm)[8]

[7] 6 H는 표준 화질 디지털 시스템 평가를 위한 설계 시청 거리(DVD)이지만, 결과가 별도로 제공된다는 조건 하에 4 H 에서의 평가자 사용도 허용된다.

[8] 디스플레이 크기가 주관적 평가 결과에 영향을 미칠 수 있다는 일부 증거가 있으므로, 실험자는 모든 실험에 사용된 디스플레이의 화면 크기와 제조사 및 모델을 명시적으로 보고하도록 요청 받는다.

8.1.4 가정 환경

이 환경은 디지털 TV 체인의 소비자 측에서 품질을 평가하는 수단을 제공하기 위한 것이다. 가정 환경에서의 SDTV 주관적 평가를 위한 특정 시청 조건은 표 8-2 에 제시된다.

표 8-2. 가정 환경에서 디지털 시스템의 주관적 평가를 위한 특정 시청 조건
Specific viewing conditions for subjective assessments of digital systems in home environment

Condition	Item	Values
a	시청 거리 대 이미지 높이의 비율	6 *H*
b	4/3 형식 비율에 대한 화면 크기	From 25" to 29"[9]
c	16/9 형식 비율에 대한 화면 크기	From 32" to 36"[10]
d	디스플레이 표준	SDTV
e	최대 휘도	200 cd/m^2
f	화면의 환경 조도(화면에 닿는 환경으로부터의 입사광은 화면에 수직으로 측정되어야 함)	200 Lux

8.1.5 평가 방법

8.1.5.1 기본 이미지 품질 평가

배포 애플리케이션용 코덱이 평가되는 경우, 이 품질은 코덱 쌍을 한 번 통과한 후 디코딩된 이미지를 의미하는 것이다. 기여 코덱의 경우, 일반적인 기여 애플리케이션을 시뮬레이션하기 위해 여러 코덱을 직렬로 연결한 후 기본 품질을 평가하는 것이다.

[9, 12] 이 화면 크기는 PVD = 6 H 에 대한 선호 시청 거리(PVD) 규칙을 만족한다.

일반적으로 텔레비전 코덱의 경우처럼 평가할 품질 범위가 좁은 경우, 사용되는 테스트 방법론은 본 권고안(Recommendation)에 기술된 DSCQS 변형 II 인 것이다. 원본 소스 시퀀스가 참조 조건으로 사용되는 것이다. 제시 시퀀스의 지속 시간에 대한 추가적인 고려가 이루어지고 있는 것이다. 최근 4:2:2 컴포넌트 비디오용 코덱에 대한 테스트에서는 본 권고안에 제시된 것과 다른 프레젠테이션으로 수정하는 것이 유리하다고 간주된 것이다. 합성 이미지는 코덱 성능을 판단할 기준이 되는 더 낮은 품질 수준을 제공하기 위해 추가적인 참조로 사용된 것이다.

평가에는 최소 6 개의 이미지 시퀀스를 사용하고, 시험 시작 전 훈련 목적으로 추가적인 시퀀스 하나를 사용하는 것이 권장되는 것이다. 시퀀스는 고려 중인 비트레이트 감소 애플리케이션의 맥락에서 중간 정도의 중요도와 높은 중요도 사이의 범위를 가져야 하는 것이다.

본 절 전반에 걸쳐, 텔레비전 비트레이트 감소의 맥락에서 중요한 이미지 시퀀스로 디지털 코덱을 테스트하는 것의 중요성이 강조된다. 따라서 특정 이미지 시퀀스가 특정 비트레이트 감소 작업에 얼마나 중요한지, 또는 한 시퀀스가 다른 시퀀스보다 더 중요한지 묻는 것은 합리적이다. 간단하지만 특별히 도움이 되지 않는 대답은 “중요도(criticality)”가 코덱마다 매우 다르다는 것을 의미하는 것이다. 예를 들어, 인트라필드 코덱의 경우 많은 디테일을 포함한 정지 이미지는 중요한 시퀀스가 될 수 있지만, 프레임 간 유사성을 활용할 수 있는 인터프레임 코덱의 경우 동일한 장면은 전혀 어려움을 주지 않을 수 있다. 움직이는 텍스처와 복잡한 움직임을 사용하는 일부 시퀀스는 모든 종류의 코덱에 중요하므로 이러한 유형의 시퀀스를 생성하거나 식별하는 것이 가장 유용하다. 복잡한 움직임은 관찰자에게는 예측 가능하지만 코딩 알고리즘에는 예측 불가능한 움직임의 형태를 띨 수 있으며, 예를 들어 복잡한 주기적 움직임이 이에 해당한다.

상관 방법, 스펙트럼 방법, 조건부 엔트로피 방법 등과 같은 이미지 중요도에 대한 가능한 통계적 척도에 대한 한 연구에서, 인트라필드/인터프레임 적응형 엔트로피 측정을 기반으로 한 간단하지만 유용한 척도가 밝혀졌다. 이 방법은 34, 45, 140 Mbit/s 코덱의 ITU-R 시험에 사용하도록 제안된 이미지 시퀀스를 ‘보정’하는 데 사용되었으며, 사용된 시퀀스의 선택에 유용함이 입증되었다. 이미지 시퀀스에 대한 이러한 측정은 이를 이미지 처리용 컴퓨터로 전송한 뒤 소프트웨어로 분석함으로써 가장 쉽게 수행된다.

이러한 기술에 접근할 수 없는 경우, 다음은 중요한 자료를 선택하는 방법에 대한 몇 가지 일반적인 지침을 제시하는 것이다.

a) 고정 단어 길이 인트라필드 코덱 (Fixed word-length intra-field codecs)

이러한 코덱들은 정지 이미지로 평가하는 것이 가능하고 타당하지만, 코딩 노이즈 과정을 더 쉽게 관찰할 수 있고 텔레비전 애플리케이션의 현실에 더 가깝기 때문에 움직이는 시퀀스의 사용이 권장된다. 정지 이미지가 코덱의 컴퓨터 시뮬레이션에 사용되는 경우, 예를 들어 소스 노이즈의 시간적 측면을 보존하기 위해 전체 평가 시퀀스에 대해 처리를 수행해야 한다. 선택된 장면은 가능한 한 많이 다음의 세부 사항들을 포함해야 한다: 정지 및 움직이는 텍스처 영역(일부는 컬러 텍스처), 다양한 방향(일부는 컬러)에서 날카롭고 높은 명암 대비 엣지(edge)를 가진 정지 및 움직이는 객체, 정적인 중간 회색 영역. 시퀀스 집합에는 최소한 하나의 시퀀스가 간신히 인지할 수 있는 소스 노이즈를 보여야 하며, 최소한 하나의 시퀀스는 합성(즉, 컴퓨터 생성)이어야 하며, 이는 스캐닝 개구(aperture)나 래그(lag)와 같은 카메라 불완전함으로부터 자유로워야 한다.

b) 고정 단어 길이 인터프레임 코덱 (Fixed word-length interframe codecs)

선택된 테스트 장면은 모두 움직임을 포함해야 하며 가능한 한 많이 다음의 세부 사항들을 포함해야 한다: 움직이는 텍스처 영역(일부는 컬러), 다양한 방향(일부는 컬러)에서 날카롭고 높은 명암 대비 엣지를 가지며 이 엣지에 수직 방향으로 움직이는 객체. 시퀀스 집합에는 최소한 하나의 시퀀스가 간신히 인지할 수 있는 소스 노이즈를 보여야 하며, 최소한 하나의 시퀀스는 합성 시퀀스여야 한다.

c) 가변 단어 길이 인트라필드 코덱 (Variable word-length intra-field codecs)

이러한 코덱들은 고정 단어 길이 코덱과 동일한 이유로 움직이는 이미지 시퀀스 자료로 테스트되는 것이 권장된다. 가변 단어 길이 코딩과 관련된 버퍼 저장 기능 덕분에, 이러한 코덱들은 이미지 전체에 걸쳐 코딩 비트 용량을 동적으로 분배할 수 있다. 예를 들어, 이미지의 절반이 많은 비트를 필요로 하지 않는 텍스처가 없는 하늘로 구성되어 있다면, 용량이 다른 부분에 절약되어 결과적으로 중요한 부분을 고품질로 재현할 수 있다. 이로부터 중요한 결론은, 이러한 코덱에 대해 이미지 시퀀스가 임계적이 되려면 화면의 모든 부분의 내용이 세부적으로 되어 있어야 한다는 것이다. 움직이는 텍스처와 정지 텍스처로 채워져 있어야 하며, 가능한 한 많은 색상 변화와 날카롭고

높은 명암 대비 엣지를 가진 객체가 포함되어야 한다. 시퀀스 집합에는 최소한 하나의 시퀀스가 간신히 인지할 수 있는 소스 노이즈를 보여야 하며, 최소한 하나의 시퀀스는 합성 시퀀스여야 한다.

d) 가변 단어 길이 인터프레임 코덱 (Variable word-length interframe codecs)

이것은 가장 정교한 종류의 코덱이며, 이를 압박하기 위해 가장 요구가 많은 자료가 필요하다. 장면의 모든 부분은 인트라필드 가변 단어 길이의 경우와 같이 세부 사항으로 채워져 있어야 할 뿐만 아니라, 이러한 세부 사항은 움직임을 나타내야 한다. 더욱이 많은 코덱들이 모션 보상 방법을 사용하기 때문에 시퀀스 전반의 움직임은 복잡해야 한다. 복잡한 움직임의 예로는 다음과 같다: 카메라의 줌과 팬이 동시에 이루어지는 장면, 바람에 펄럭이는 텍스처 또는 세부적인 커튼이 배경에 있는 장면, 3 차원 공간에서 객체가 회전하는 장면, 세부적인 객체가 화면을 가로질러 가속하는 장면. 모든 장면은 서로 다른 속도, 텍스처 및 높은 명암 대비 엣지를 가진 객체의 상당한 움직임과 다양한 색상 내용을 포함해야 한다. 시퀀스 집합에는 최소한 하나의 시퀀스가 간신히 인지할 수 있는 소스 노이즈를 보여야 하며, 최소한 하나의 시퀀스는 자연 정지 이미지로부터 복잡한 컴퓨터 생성 카메라 움직임을 가져야 하며(따라서 노이즈와 카메라 래그가 없음), 최소한 하나의 시퀀스는 완전히 컴퓨터로 생성되어야 한다.

8.1.5.2 후단 처리 이후의 화질 평가

배포 애플리케이션용 코덱이 평가되는 경우, 이 품질은 코덱 쌍을 한 번 통과한 후 디코딩된 이미지를 의미하는 것이다. 기여 코덱의 경우, 일반적인 기여 애플리케이션을 시뮬레이션하기 위해 여러 코덱을 직렬로 연결한 후 기본 품질을 평가하는 것이다. 이 평가는 특정 후처리 과정(예: 컬러 매트, 슬로 모션, 전자 줌 등)에 대해 코덱이 기여 애플리케이션에 적합한지를 판단할 수 있도록 하기 위한 것이다. 이러한 평가를 위한 최소 장비 구성은 시험 대상 코덱을 한 번 통과한 후, 관심 있는 후처리 과정을 거치고, 그 다음 시청자에게 제시하는 방식이다. 그러나 후처리 이후에 추가적인 코덱을 사용하는 것이 기여 애플리케이션을 더 잘 대표할 수 있다.

사용되는 테스트 방법론은 DSCQS의 변형 II(Variant II)이다. 그러나 이 경우에는 참조 조건이 디코딩된 이미지와 동일한 후처리를 거친 소스가 된다. 만약 낮은 품질의 참조

영상을 포함하는 것이 유리하다고 판단된다면, 그 참조 영상 또한 동일한 후처리를 거쳐야 한다.

후처리 평가에 필요한 테스트 시퀀스는 다른 디지털 애플리케이션용 시퀀스와 정확히 동일한 중요도 기준(criticality criteria)을 따른다. 그러나 크로마 키(Chroma key) 전경 시퀀스의 경우, 일반적으로 세부 요소가 없는 파란색 배경이 상당 부분을 차지하기 때문에 이러한 기준을 충족시키는 것이 어려울 수 있다.

실제로 여러 후처리 과정에 대해 코덱을 평가해야 할 수도 있는 실용적 제약 때문에, 사용되는 테스트 이미지 시퀀스의 수는 최소 3 개일 수 있으며, 추가로 1 개를 시연(demonstration) 용도로 사용할 수 있다. 시퀀스의 특성은 평가 대상인 후처리 작업에 따라 달라질 수 있지만, 텔레비전 비트레이트 감소 및 해당 후처리 과정의 맥락에서 중간 수준의 중요도에서 높은 중요도에 이르는 범위여야 한다. 슬로 모션 평가의 경우 소스 재생 속도의 1/10 수준의 디스플레이 속도가 적합할 수 있다.

8.1.5.3 열화 특성 평가

코덱 영상에서 전송 또는 송출 채널의 결함으로 인해 발생하는 손상에 대한 주관적 평가에서는, 최소 5 개 이상의(가능하다면 더 많은) 비트 오류율 또는 선택된 전송/송출 조건을 선택해야 하며, 이는 대략적으로 로그 간격으로 배치되어야 하고, 코덱 손상이 “인지할 수 없음”에서 “매우 거슬림”에 이르는 범위를 충분히 샘플링해야 한다.

코덱 평가가 매우 낮은 전송 비트 오류율에서 요구될 수 있으며, 이 경우 10 초 테스트 시퀀스 동안에는 거의 발생하지 않을 정도로 드문 가시적 전이(transient)가 나타날 수 있다. 이 경우 여기에서 제안된 제시 타이밍은 명백히 이러한 테스트에 적합하지 않다.

만약 낮은 비트 오류율 조건에서(10 초 동안 소수의 가시적 전이만 발생하는 조건) 코덱 출력을 녹화하여 이후 주관적 평가 제시용으로 편집하려는 경우, 사용되는 녹화 영상이 장시간 동안 코덱 출력을 시청했을 때의 전형적인 상태를 잘 반영하도록 주의해야 한다.

코덱 성능을 전송 비트 오류율의 여러 범위에 걸쳐 탐색할 필요가 있기 때문에, 실용적 제약을 고려하면 3 개의 테스트 이미지 시퀀스와 추가적인 1 개의 시연용 시퀀스가 적절할 것으로 보인다. 시퀀스 길이는 약 10 초 정도가 적당하지만, 테스트

시청자들은 15 초에서 30 초 정도의 길이를 선호할 수 있음에 유의해야 한다. 시퀀스는 텔레비전 비트레이트 감소 맥락에서 중간 수준에서 높은 수준의 중요도를 가져야 한다.

테스트는 손상의 전체 범위를 포괄하므로, DSIS 이 적절하며 이를 사용해야 한다.

8.1.5.4 이미지 콘텐츠 열화 특성

표준 화질 디지털 텔레비전 시스템에 적용하려면 다음 절차를 사용해야 한다.

8.1.5.4.1 중요도(Criticality)의 정의

시험 중인 디지털 텔레비전 시스템의 특성을 나타내며 객관적 측정에 의해 측정되는 '중요도(criticality)'라고 불리는 특정한 척도를 정의해야 한다. 디지털 텔레비전 시스템의 예로 MPEG-2 MP@ML 이 사용되며, ITU-R BT.1210 에 기술된 엔트로피 기반 중요도 측정의 고정 양자화기(fixed quantizer) 방법이 적용된다.

8.1.5.4.2 영상 콘텐츠 오류 특성 도출 절차

시험 중인 디지털 텔레비전 시스템의 특성을 나타내며 객관적 측정에 의해 측정되는 '중요도(criticality)'라고 불리는 특정한 척도를 정의해야 한다. 디지털 텔레비전 시스템의 예로 MPEG-2 MP@ML 이 사용되며, ITU-R BT.1210 에 기술된 엔트로피 기반 중요도 측정의 고정 양자화기(fixed quantizer) 방법이 적용된다.

- 1 단계: 주관적 평가에 사용되는 테스트 시퀀스의 중요도 측정

아래 3 단계에서 설명되는 주관적 평가에 사용되는 테스트 시퀀스의 중요도를 측정한다. 그림 8-1 은 예시 시스템에 대해 각 시퀀스의 평균값과 표준편차를 보여준다. 대부분의 시퀀스는 0.8~1.4 bits/pixel 범위의 중요도 값을 가진다. 일부 시퀀스는 영상 콘텐츠가 시퀀스 동안 크게 변하기 때문에 표준편차가 큰 값을 가진다.

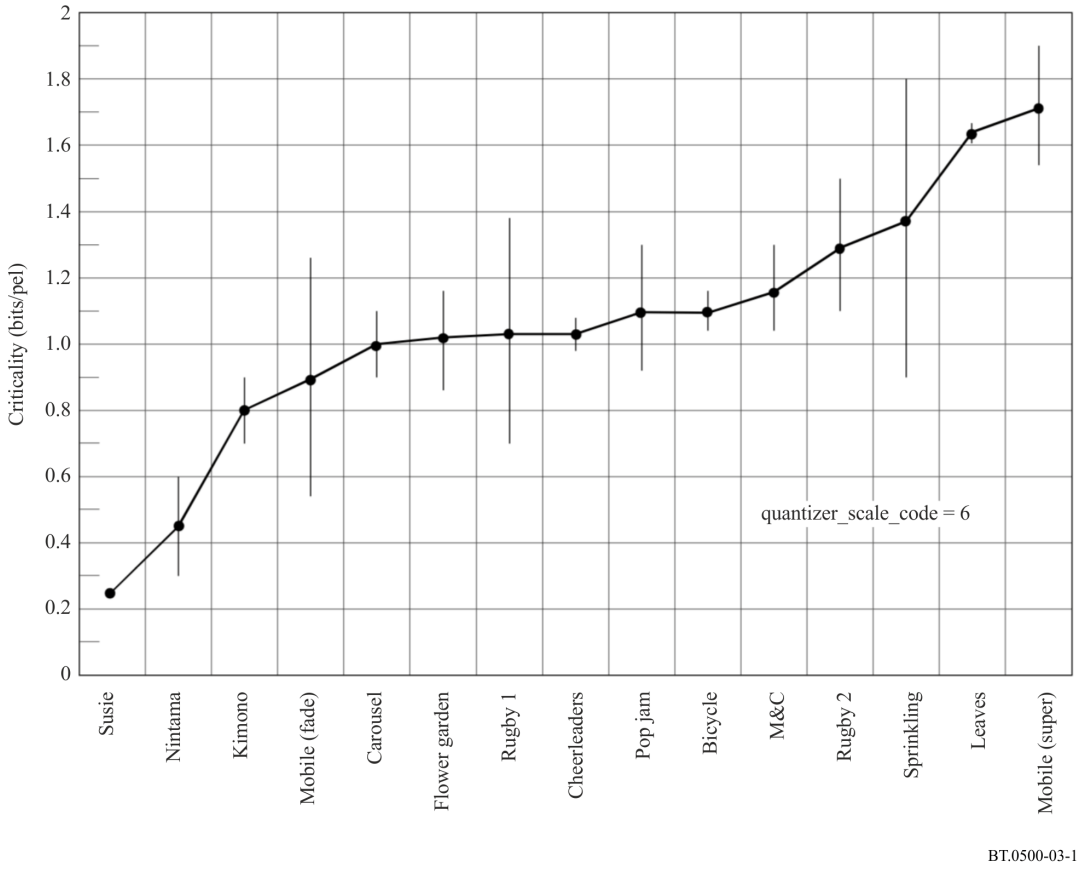


그림 8-1. 테스트 시퀀스 중요도의 평균값과 표준편차

Means and standard deviation of criticality of test sequences

- 2 단계: 장기간에 걸친 방송 프로그램의 중요도 분포 측정

방송 텔레비전 프로그램의 중요도 분포를 충분히 긴 기간(예: 1 주일) 동안 측정한다. 그림 8-2 는 NTSC 방송 신호를 측정용 컴포넌트 Y/C 신호로 변환하여 1 주일 동안 총 130 시간에 대해 측정한 분포의 예를 보여준다. 텔레비전 프로그램의 중요도 발생 빈도는 5×10^{-3} bits/pixel 간격으로 계산되었다. 이 그림에는 주관적 평가에 사용된 테스트 시퀀스의 중요도도 함께 표시되어 있다.

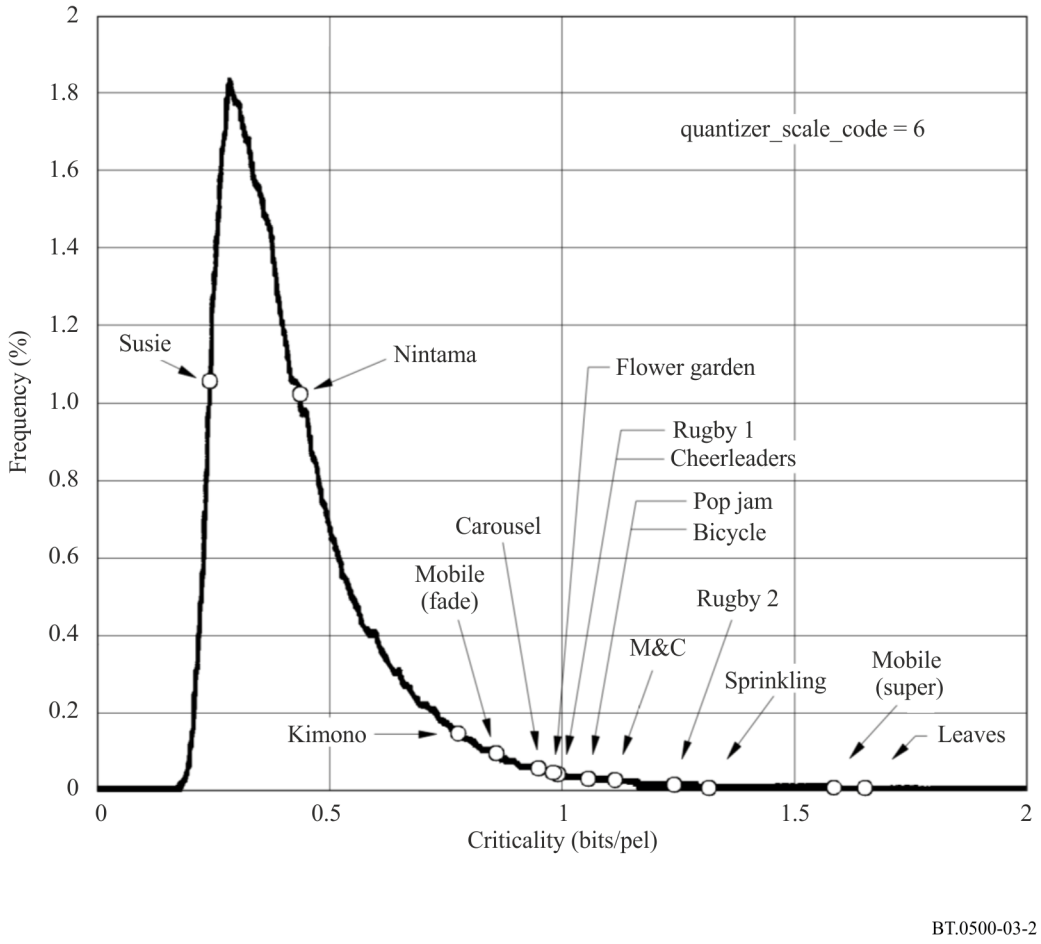


그림 8-2. 방송 프로그램의 중요도 분포와 테스트 시퀀스의 중요도

Distribution of criticality for broadcast programmes and criticality of test sequences

- 3 단계: 시험 대상 시스템의 화질에 대한 주관적 평가를 수행하고, 중요도와 주관적 화질 간의 관계를 도출한다.

디지털 텔레비전 시스템의 화질은 DSCQS 을 사용하여 평가한다. 1 단계에서 얻은 중요도와 주관적 평가 결과를 결합하여, 중요도와 평가 점수 간의 관계를 도출한다. 그림 8-3 은 예시 시스템의 4, 6, 9, 15 Mbit/s 비트레이트에서의 화질을 보여준다. 그림에 나타난 품질 차이(DSCQS %)는 기준인 원본 4:2:2 컴포넌트 시퀀스로부터의 열화를 나타낸다. 그림 3-4 는 중요도와 품질 차이의 관계를 보여준다. 이 예시에서는 중요도와 화질 사이에 선형 관계가 있다고 가정하고, 최소자승법을 사용하여 회귀선을 도출하였다. 각 비트레이트에 대한 회귀선이 그림에 나타나 있다. 일반적으로 주관적 평가 결과에 따라 비선형 관계를 적용할 수도 있다.

- 4 단계: 3 단계(중요도 대 화질)와 2 단계(중요도 대 발생 빈도)의 결과를 결합하여 영상 콘텐츠 오류 특성(화질 대 발생 빈도)을 도출한다.

2 단계와 3 단계에서 얻은 결과를 결합하여, 즉 디지털로 부호화된 텔레비전 프로그램의 화질 분포를 통해 영상 콘텐츠 오류 특성을 도출한다. 방송 텔레비전 프로그램의 화질 열화는 누적 발생 빈도로 변환된다. 그림 8-5 는 예시 시스템의 영상 콘텐츠 오류 특성을 보여준다.

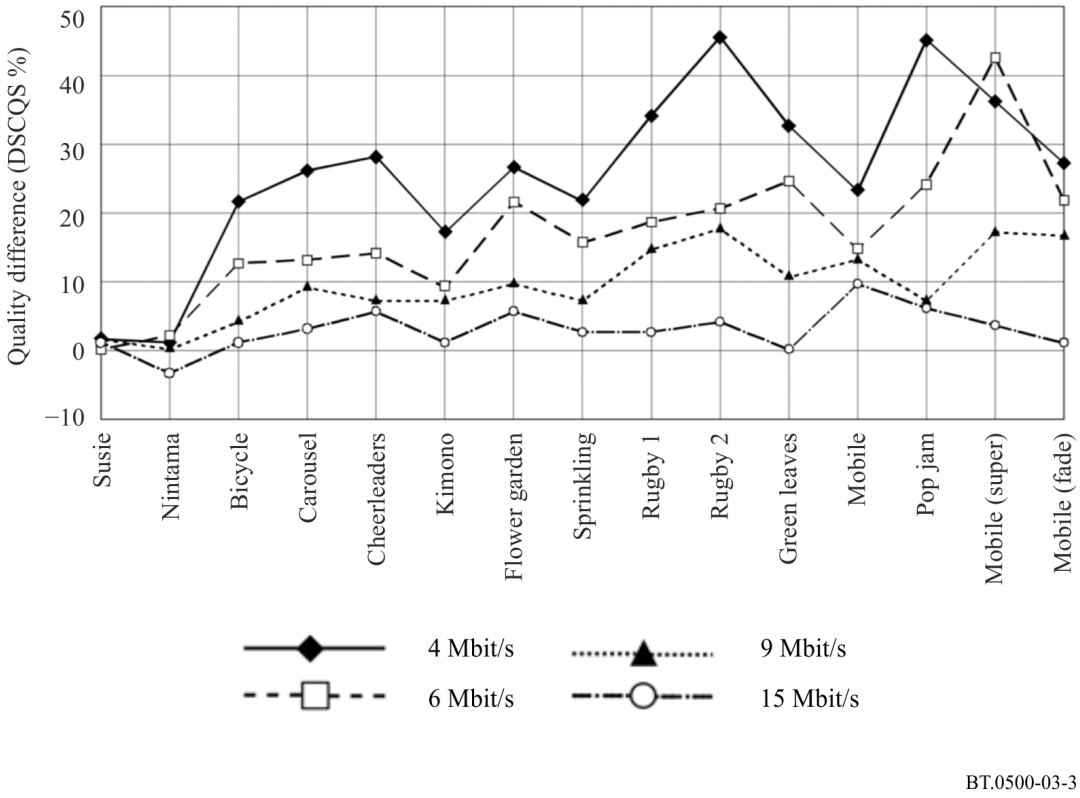


그림 8-3. 주관적 평가 결과 (MP@ML, 6H)

Results of subjective assessment (MP@ML at 6H)

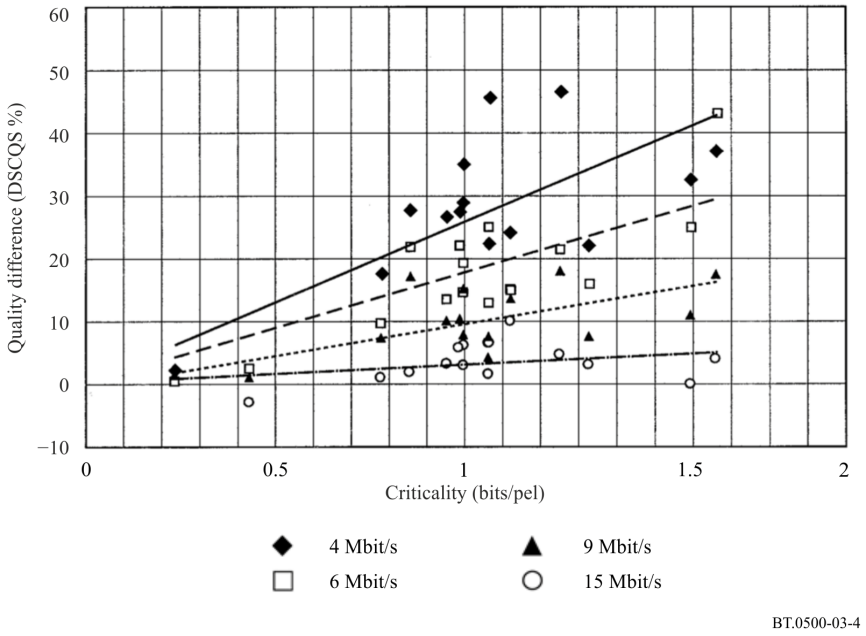


그림 8-4. 중요도와 평가 점수의 관계 (MP@ML, 6H)

Relationship between criticality and assessment score (MP@ML at 6H)

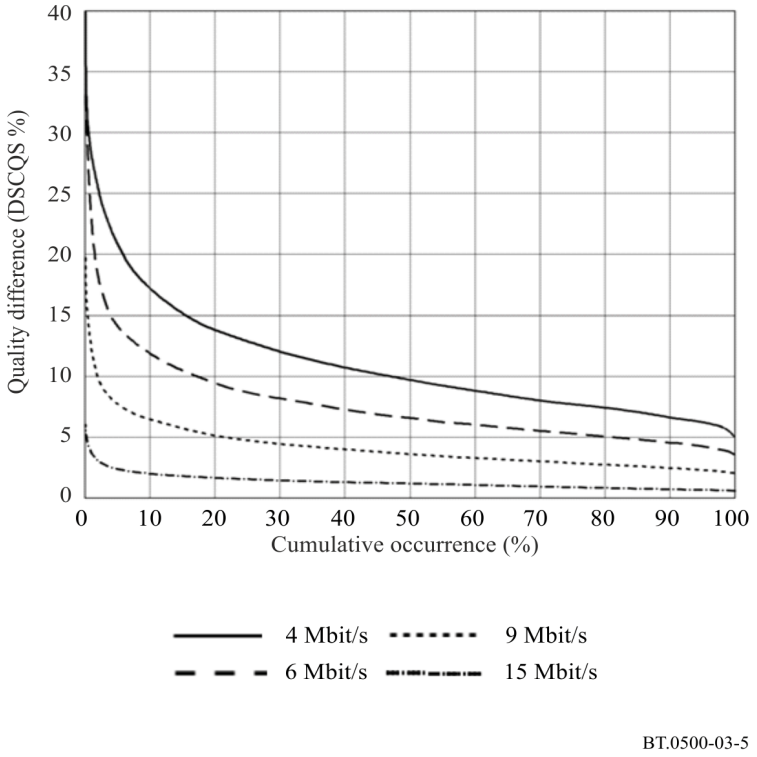


그림 8-5. 영상 열화의 누적 발생 빈도 (MP@ML, 6H)

Cumulative frequency of occurrence of image degradation (MP@ML at 6H)

8.1.6 적용 참고사항

절대적인 코덱 품질 또는 손상에 대한 판단이 필요하지 않고 단순히 순위만 필요하거나, 또는 이중 자극 결과에서 얻은 순위를 확인하고자 하는 경우, 쌍자극 비교(paired-stimulus comparison) 방법을 사용해야 한다.

이 권고안에서 설명된 바와 같이, 이 방법은 민감한 비교를 가능하게 하며, 시스템 쌍 간의 관계를 측정할 수 있는 수단을 제공한다. 이 방법은 두 개 이상의 시스템의 품질 또는 손상에 대한 순위를 매기도록 확장할 수 있다. 이러한 접근 방식에서는 관찰자들이 모든 가능한 이미지 시퀀스 쌍에 대해 순위를 매긴 결과로부터 전체 순위가 도출된다.

분석은 관찰자가 예를 들어 이미지 A를 B보다, B를 C보다, 그리고 C를 다시 A보다 더 낫다고 평가할 수 있다는 사실로 인해 복잡해진다. 이러한 경우를 “비이행 삼항(intransitive triad)”이라고 한다.

이 방법의 문제점은 테스트 이미지 시퀀스 및 코덱의 수가 늘어날수록 제시해야 하는 횟수가 제곱 비율로 증가하여 비현실적일 수 있다는 점이다.

방송 채널이 다중 프로그램 스트림 또는 스케일러블/계층형 부호화 방식을 통해 전송되는 경우, 다음 사항들을 고려하여 평가 방법을 조정할 필요가 있다.

수용 가능한 서비스의 기준이 소스 코딩에서의 완전한 투명성(transparency)이 아닐 수도 있다. 대신 주어진 비트레이트 할당에서 시스템이 기존 서비스의 실질적인 대안으로 기능할 수 있는 능력이 기준이 될 수 있다. 이에 따라 품질 테스트에서 참조 자료로는 비압축 디지털 형태의 원본이 아니라 일반적인 수신 조건에서 기존 시스템을 통해 전달되는 자료를 사용하는 것이 적절할 수 있다. 또한 현재 및 미래의 프로그램 콘텐츠 범위를 대표하는 테스트 자료를 사용하는 것도 적절할 수 있다. 테스트 시 일반적인 테스트 방법은 DSCQS을 사용해야 한다.

시스템이 전체 채널 부하 및 전송 손상 조건에서 개별 프로그램 스트림의 무결성을 유지할 수 있는 능력이 중요한 문제이다. 이에 따라 손상 테스트에서는 전체 채널 부하를 보장하고, 다양한 수신 조건 범위를 대표하는 손상 수준을 사용하는 것이 적절할 수 있다. 테스트 시 일반적인 테스트 방법은 DSIS을 사용해야 한다.

참고 - 아날로그와 디지털 시스템을 동일한 맥락에서 평가하는 경우, 아날로그와 디지털 시스템 모두에 대해 난이도가 균형 있게 반영되는 테스트 자료 집합을 선택하는

것이 중요하다. 이러한 경우 보조적인 분석을 위해 다차원 척도법(multidimensional scaling) 절차를 적용하는 것이 유용할 수 있다.

8.2 HDTV

8.2.1 시청 환경

아래 표 8-3 에 명시되지 않은 경우, 시청 환경은 6 장에 기술된 바와 같아야 한다.

표 8-3. HDTV 이미지 품질의 주관적 평가를 위한 시청 조건
Viewing conditions for the subjective assessment of HDTV image quality

조건	항목	값
a	시청 거리 대 이미지 높이의 비율	3
b	화면의 최대 휘도 (cd/m^2)[11]	150-250
c	비활성 화면의 휘도 대 최대 휘도의 비율[12]	≤ 0.02
d	완전히 어두운 방에서 블랙 레발만 표시할 때의 화면 휘도 대 최대 화이트에 해당하는 휘도의 비율[13]	약 0.01
e	이미지 디스플레이 뒤 배경의 휘도 대 이미지 최대 휘도의 비율	약View 0.15
f	다른 광원으로부터의 조명[14]	low
g	배경의 색도	D_{65}

[11] 100% 진폭의 비디오 신호에 해당하는 화면의 최대 휘도.

[12] 이 항목은 디스플레이의 명암비 범위뿐만 아니라 실내 조명의 영향을 받을 수 있다.

[13] 블랙 레벨은 0% 진폭의 비디오 신호에 해당한다.

[14] 실내 조명은 조건 c 와 e 를 만족시킬 수 있도록 조정되어야 한다.

h	위 사양을 만족하는 배경 부분에 의해 형성되는 각도[15]. 이는 모든 관찰자에게 유지되어야 함	53° high × 83° wide
i	관찰자의 배치	디스플레이 중앙으로부터 수평으로 ±30° 이내. 수직 한계는 연구 중
j	디스플레이 크기[16]	1.4 m (55 in)

8.2.2 평가 방법

방송 시스템에 의해 전달되는 HDTV 이미지의 전반적인 품질에 대한 주관적 평가는 HDTV 스튜디오 품질 이미지를 참조로 하여 DSCQS (7.2 절)을 사용하여 이루어져야 한다.

HDTV 방송 시스템의 열화 특성 평가는 HDTV 스튜디오 이미지 또는 손상되지 않은 방송 이미지를 참조로 하여 DSIS 방법(7.1 절)을 사용하여 이루어져야 한다.

이러한 방법을 사용할 때, 디스플레이 형식의 영향과 기본 시스템 형식(예: 업컨버전)의 영향을 구별하도록 주의해야 한다. 적용 가능하고 적절하다고 판단되는 경우, 디스플레이 형식의 차이를 고려하여 다른 디스플레이를 사용하여 보충 평가를 수행할 수 있다.

일부 HDTV 방송 시스템은 내장된 기존 텔레비전 형식(하위 호환성)을 포함할 수 있다. 따라서 이미지 품질 측면에서 HDTV 방송에 내장된 기존 텔레비전 이미지의 적절성을 평가할 필요가 있다. 이러한 시스템의 경우, 8.1 절에 제시된 시청 조건 및 평가 방법을 적용해야 한다.

8.1 절에 기술된 기본 개념 및 절차는 비트레이트 감소 방식을 사용하는 디지털 HDTV 방송 시스템에 적용되어야 한다.

---

[15] 최소 높이 28° × 너비 48°가 권장된다.

[16] 지정된 크기의 디스플레이를 사용할 수 없는 경우 ³ 76.2 cm (30 in) 값을 사용해야 한다.

8.2.3 테스트 자료

보고서 ITU-R BT.2245 는 광범위한 정지 이미지와 동영상 시퀀스를 나열하고 있다. 이들은 HDTV 품질 평가를 위한 공통 테스트 자료로 사용하는 것이 바람직하다.

8.3 Multi-programme service

다중 프로그램 서비스 내에서 CBR 로 압축 및 코딩된 개별 프로그램의 품질에 대한 주관적 평가를 위해, 8.1 절 또는 8.2 절에 상세히 기술된 주관적 절차와 8.3.2 절에 기술된 절차를 사용해야 한다.

다중 프로그램 서비스 내에서 통계적 다중화 또는 공동 코딩과 같은 방법을 사용하여 VBR 로 압축 및 코딩된 개별 프로그램의 품질에 대한 주관적 평가를 위해, 8.1 절 또는 8.2 절에 상세히 기술된 주관적 절차와 이 부록의 8.3.3 절에 기술된 절차를 사용해야 한다.

8.3.1 일반 평가 세부사항

- 주제 기반 채널의 품질 평가는 해당 채널에서 일반적으로 전송되는 것과 유사한 콘텐츠 및 중요도를 가진 테스트 자료를 사용하여 수행되어야 한다.

- 시간 경과에 따라 '순간적인' 품질이 변하는 프로그래밍의 전반적인 인지 품질을 평가하기 위해, 8.3.2 절과 8.3.3 절에 기술된 절차를 사용해야 한다.

- DSCQS 설명에 포함된 의견에 따라, 낮은 품질의 참조를 포함하는 시스템에 대한 결과의 스케일링은 다중 프로그램 서비스를 낮은 품질의 자료와 비교하는 테스트에 적용되고 추가적으로 연구되어야 한다.

8.3.2 CBR 다중 프로그램 서비스에 대한 주관적 이미지 평가 절차

각 SDTV 및 HDTV 프로그램에 대한 주관적 이미지 품질 평가는 8.1 절(SDTV) 또는 8.2 절(HDTV)에 기술된 방법을 사용하여 독립적으로 수행될 수 있다. 시스템 기본 품질 평가를 위해서는 일반 테스트 방법인 DSCQS 를 사용해야 한다. 전송 손상이 있는 프로그램의 평가를 위해서는 일반 테스트 방법인 DSIS 를 사용해야 한다.

8.3.3 VBR 다중 프로그램 서비스에 대한 주관적 이미지 품질 평가 절차

VBR 로 인코딩된 SDTV 및 HDTV 프로그램의 주관적 이미지 품질 평가는 DSCQS 을 사용하여 수행될 수 있다. 이미지 품질은 다중화된 모든 프로그램의 이미지 콘텐츠에 따라 달라질 수 있으므로, 테스트 자료의 선택에도 주의를 기울여야 한다.

9 결론

본 보고서는 국제전기통신연합 무선통신부문(ITU-R)의 BT.500 권고안을 바탕으로, 주관적 영상 화질 평가의 기본 개념, 실험 절차, 통계적 분석 방법, 그리고 응용 분야별 평가 기법을 종합적으로 분석하였다. 이를 통해 UHDTV 환경에서 요구되는 새로운 화질 평가 체계의 기초를 마련하고, 향후 UHDTV 서비스의 품질 평가 방법론 수립 시 국제표준과의 정합성을 확보하기 위한 기술적 근거 자료를 제공한다.

본 분석 결과는 향후 UHDTV 서비스에 적합한 화질 평가 기준과 절차를 구체화하고, 이를 기반으로 표준화 추진 및 품질 측정 자동화 기술 개발에 활용될 수 있을 것으로 기대된다. 또한 본 연구에서 제시한 분석 내용은 향후 국내외 방송·미디어 산업의 UHDTV QoE 및 국제 표준화 활동 추진 시 실질적인 참조자료로 활용될 것으로 기대된다.

부 록 I-1

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

참고 문헌

[1] ITU-R Recommendation BT.500-14, Methodologies for the Subjective Assessment of the Quality of Television Pictures, International Telecommunication Union - Radiocommunication Sector, Geneva, Switzerland, 2023. (국제전기통신연합 무선통신부문, 「텔레비전 영상 품질의 주관적 평가 방법」, 2023.)

부 록 II-2

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

기술보고서의 이력

판수	채택일	표준번호	내용	담당 위원회
제 1 판	2025.12.05	제정 FBMF-STD-xxx	ITU-R 에서 제정한 BT.500 권고안의 주요 내용을 심층적으로 분석하여, 영상 화질에 대한 주관적 품질 평가 절차, 관찰자 선별 기준, 통계적 분석 방법, 및 신뢰구간 산출 기법 등을 고찰	UHD 융합기술 분과위원회