ELEKTROTEHNIŠKI VESTNIK 79(4): 165-168, 2012
ERK 2012 CONFERENCE ISSUE

# Koncept več-modalnega evalvatorja kakovosti multimedije

**Tomaž Lovrenčič[1], Mitja Štular[1], Andrej Žgank[2]**

[1]*Telekom Slovenije d.d., Cigaletova 15, 1000 Ljubljana, Slovenija*
[2] *Univerza v Mariboru, Inštitut za elektroniko in telekomunikacije, Smetanova ul. 17, 2000 Maribor, Slovenija*
*e-pošta: tomaz.lovrencic@telekom.si*

## A concept of multi-modal evaluation of multimedia quality

*The paper presents a concept of multi-modal evaluation of quality for measuring high-definition multimedia degraded from a network or other impairments. Sophisticated audio and video metric algorithms enhanced with a structure detector are implemented. The detector reveals the Regions Of Interest known to be importantly and impacting the overall quality score in case of degradation. This evaluator is a useful tool in investigating the interaction between the different modal streams perceived by the end-user.*

## 1 Uvod

Vrednotenje kakovosti storitve je ključno za ohranjanje zadovoljstva uporabnikov in s tem obstoja ter razvoja konkurenčne ponudbe več-modalnih vsebin na tržišču. Kakovost storitve se v realni aplikaciji spreminja in je, po navadi, funkcija elementov na prenosni poti na katere ponudnik storitve nima vpliva. Primeri takšnega vpliva na prenosni poti so izguba in zakasnitev IP paketov, transkodiranje ter kompresijske deformacije podatkov. V teh primerih želi ponudnik ovrednotiti stopnjo degradacije, ki ji je storitev podvržena z namenom, da lahko v nadaljevanju ukrepa in tako zagotovi pričakovani nivo storitve. To je še posebej pomembno pri vsebinah visoke ločljivosti, kjer so pričakovanja uporabnikov večja: povečevanje ločljivosti videa (HD Ready, Full HD in več), pasovne širine avdia (wide-band, superwide-band), želene višje hitrosti prikaza okvirjev (fps), manjše število video- in avdio-napak. Sistem vrednotenja, ki spremlja kakovostni nivo, imenujemo evalvator kakovosti storitve. Evalvator načeloma vsebuje eno ali več objektivnih metrik, ki na podlagi postavitve sistema in uporabe daje kvalitativno ali kvantitativno oceno algoritma, ki jo lahko, s primerno preslikavo, primerjamo z vrednostjo srednje povprečne ocene (MOS), dobljene na osnovi subjektivnega vrednotenja. Tradicionalne metrike, kot npr. maksimalno razmerje signal-šum (PSNR) delujejo zadovoljivo in procesorsko hitro, vendar zaradi narave dojemanja okolja človek sprejema dražljaje v obliki struktur in ne na osnovi pikslov, kot to predpostavlja PSNR, kadar govorimo o vizualni informaciji. Napredni algoritmi kot je na primer indeks strukturne podobnosti (SSIM), omogočajo prav to. Podobno v avdio-modalnih vsebinah psiho-akustični model, kot je definiran v metriki ocene kakovosti glasu (PESQ), oceni kognitivno pomembnost intervala glasu v posnetku. Čeprav sofisticirani pristopi vrednotenja kakovosti posamičnih modalnosti dobro korelirajo z MOS, je potrebno dodatno pozornost usmeriti v izkušnjo s storitvijo iz perspektive končnega uporabnika. Več-modalne vsebine namreč nanj vplivajo z dražljaji, ki delujejo na človeški senzorni sistem s prepletenim učinkom, npr. korelacija slišno-vidnega signala. Zato je tudi pri izbiri in izdelavi sistema vrednotenja kakovosti storitve smiselno upoštevati križno-modalno interakcijo [1]. Kakovost uporabniške izkušnje (QoE) med drugim obravnava tudi osredotočenost uporabnika na območja interesa oz. ROI (angl. Region Of Interest). Sistem osredotočenosti na ROI vsebuje detektor polj ROI, ki v primeru zvočne vsebine razloči »iskano« avdio ovojnico (npr. glas) iz množice vseh frekvenc v frekvenčnem spektru (glas+nepomembna informacija) ali se, v primeru video vsebine, vizualno osredotoči na polje pikslov, ki zajemajo obraz napovedovalca in ne na ozadje. Pomembnost ROI pri vrednotenju kakovosti storitve je zato precej večja od ostalih območij, zato so tudi napake v tem predelu, gledano iz perspektive uporabnika, manj zaželene [2].

V tem prispevku predlagamo integracijo video in avdio kakovostne metrike ter detektorja polj ROI v skupen programski paket - multimedijski evalvator, ki omogoča nadaljnje raziskovalno delo na področju optimiziranja algoritmov za objektivno evalvacijo multimedije visoke kakovosti in križno-modalnih učinkov. Za namene verifikacije delovanja evalvatorja smo uporabili testni material iz komercialne produkcije s poudarkom na kontekstu tipa novice in reportaže. Vrednotenje kakovosti avdio dela multimedije smo izvedli z algoritmom za evalvacijo posnetkov človeškega glasu PESQ, kakovost videa pa z algoritmom SSIM z dodatnim detektorjem polj ROI. Primerjalno smo opravili še vizualno evalvacijo kakovosti s PSNR, kar je omogočilo delovanje evalvatorja v realnem času na testnem sistemu.

## 2 Vrednotenje kakovosti storitve

Prenos multimedijske vsebine je zaradi velike količine podatkov in narave uporabe v veliki meri občutljiv na razne variacije pri distribuciji le-tega. Prenosni kanal realno-časovnih aplikacij, povečini temelječ na transportnem protokolu UDP, ne zagotavlja 100% zanesljivosti prenosa, kar še posebej velja v heterogenih okoljih (internet). Nepravilnosti se pojavijo predvsem

zaradi izgube, latence in časovne variance prihoda UDP paketov ter zaradi asinhronizacije različnih kanalov (avdio-video ali video-video pri kodiranju »Multi View Coding«). Sledenje sekvenčnim številkam paketov in detekcijo izgubljenih paketov višje-nivojski protokol, npr. RTP (angl. Real Time Protocol), ki pa je omejen s predefiniranimi QoS postavkami, kot so količina sprejemnega pomnilnika, detekcija in korekcija napak ter stopnja ponovnega pošiljanja. Omenjene omejitve vplivajo na zaznavno kakovost multimedijskega toka, kot ga vidi uporabnik in od tega je odvisno zadovoljstvo s storitvijo, zaznano s strani uporabnika. Dodatno je pri tem potrebno upoštevati parametre, ki niso neposredno odvisni samo od napak v omrežju. Nekateri takšni parametri so različna pomembnost RTP paketov, velikost in ločljivost zaslona predvajalne naprave, itd.

Tradicionalne metode določanja kakovosti videa temeljijo na subjektivnem testiranju z naborom testnih scenarijev, kar poteka neposredno na reprezentativnem vzorcu uporabnikov. S tem dobimo natančno uporabniško oceno kakovosti storitve - MOS. Subjektivni testi kakovosti storitve imajo tudi svoje slabosti, npr. lahko so finančno zahtevni ter časovno zamudni. Pristop z računalniško implementiranimi objektivnimi metrikami premosti omenjene težave. Primer takšnega pristopa je izračun PSNR iz povprečja kvadrata napake (MSE) med referenčnim, ter testnim signalom na različnih kanalih modalnosti, načeloma ločeno za avdio in video vsebine, kot prikazujeta enačbi (1) in (2) za video vsebino. Pri tem je *M*N* velikost referenčne ($I_1$) in testne ($I_2$) slike v trenutku *t* ter *R* maksimalna absolutna razlika med 2 vzorcema (za 8-bitno sliko je ta vrednost *255*).

$$MSE = \frac{\sum_{M,N}[I_1(m,n) - I_2(m,n)]^2}{M * N} \quad (1)$$

$$PSNR = 10log_{10}\left(\frac{R^2}{MSE}\right) \quad (2)$$

Izračun PSNR je sicer računsko nezahteven , vendar študije kažejo v določenih primerih uporabe (razporejen Gaussov šum na celotni sliki) nesimilarnost kakovostne ocene pridobljene s PSNR s tisto, dobljeno iz subjektivnih testov [3]. Druge verzije kakovostnih metrik zato vključujejo reproduciranje percepcijskega modela človeškega vida oziroma sluha (angl. HVS-Human Visual System model, HAS- Human Auditory System model) z informacijo o pomembnosti intenzitete in ranga dražljaja ter zakrivanju oz. maskiranju le-tega. Dokazano daje HVS večjo pomembnost *zaznavi spremembe strukturne informacije* in ne samo *zaznavi napak* [4]. Indeks strukturne podobnosti (SSIM) je tako primerna metoda za merjenje podobnosti dveh slik, kot prikazuje enačba (3). Pri tem sta *x* in *y* primerjani sliki oz. polja pikslov, $\mu_x$ in $\mu_y$ povprečje *x* in *y*, $\sigma_x^2$, $\sigma_y^2$ varianci *x* in *y*, $\sigma_{xy}$ kovarianca med slikama ter $c_1$ in $c_2$ konstanti, ki stabilizirata deljenje s šibkim imenovalcem [5].

$$SSIM(x,y) = \frac{(2 + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (3)$$

Analogno v avdio svetu tradicionalne metode vrednotenja kakovosti storitve obravnavajo meritve razmerja signal-šum in delež popačenja, ki pa jih vse pogosteje nadomeščajo objektivne analize s psiho-akustičnim modelom, kot sta na primer metodi Perceptual Evaluation of Audio Quality (PEAQ) [6] in Perceptual Evaluation of Speech Quality (PESQ) [7]. Omenjeni metodi se razlikujeta v tipu avdio izvora, ki ga vrednotita, pri tem slednji implementira 2 različici: ozkopasovno (PESQ-NB), za glasovne signale med 300-3400 Hz ter širokopasovno za avdio sisteme s pasovno širino signala najmanj med 50-7000 Hz (PESQ-WB). Opisani algoritmi so metrike s polno referenco (FR metrike), ki veljajo za najprimernejšo izbiro, kadar želimo natančno oceno kakovosti med referenčnim signalom in testnim oz. degradiranim signalom. Ker FR metrike zahtevajo dostop do referenčnega gradiva, je uporaba povečini omejena na testne raziskave v nadzorovanem okolju. V praksi še poznamo metode z a) delno referenco (RR metrike) in b) brez reference (NR metrike), ki metriki priskrbijo le a) določene parametre ali pa so celo b) brez informacij o originalu.

Koncepti vrednotenja kakovosti multimedije zajemajo tudi križno-modalno interakcijo in s tem bolj celovit približek izračunov z oceno MOS. Dokazano je [8], da se skupna kakovost (angl. overall quality) poslabša tudi v primeru, ko video in avdio del multimedije ne tvorita sinhroniziranega, koherentnega podatkovnega toka. Sodobni trendi uvajajo združeno rešitev, tj. multimedijski evalvator, ki zmore delovati v več-modalnem načinu in je tako primernejše za več-modalne vsebine [9].

## 3 Polja ROI v videu in detekcija obraza ter ust

Polje ROI je splošno ime za območje na sliki z iskanimi lastnostmi. Na nivoju pikslov govorimo o poljih z večjo osvetljenostjo, poljih močnega kontrastnega gradienta (robovi) in podobno. Na višjem programskem nivoju pa vizualna obdelava podatkov omogoča napredne funkcije uporabe, kot npr. detekcija objektov na sliki, tj. človeškega obraza in ust. Ker gre pri tem za človeku lastno pojmovanje, je potrebno detektorju omogočiti razumevanje posameznega pojma. Temu postopku rečemo strojno učenje. Prva faza učenja vsebuje treniranje detektorja na setu referenčnih vzorcev, kjer se algoritem detektorja uči avtomatične razpoznave kompleksnih vzorcev in razumevanja vhodnih podatkov. Težavnost tega procesa predstavlja dejstvo, da je kombinacij vseh možnih vhodnih stanj preveliko, da bi jih lahko nedvoumno opisali z množico referenčnih vzorcev in trivialno povezali z možnimi izhodnimi stanji detektorja. Naloga algoritma je zato generalizacija vedenjskih vzorcev na podatkih za treniranje tako, da v primeru drugih testnih podatkov daje pričakovan odziv. S postopkom učenja dobimo

učni model, npr. klasifikator, kateri določuje vedenjske vzorce in odziv na neznanem testnem gradivu.

Učenje modelov poteka na bazi pozitivnih in negativnih vzorcev, ki učnemu algoritmu predložita podatke o iskanem objektu. Primer takšnega strojnega učenja je algoritem AdaBoost [10]. AdaBoost v času učenja modela adaptivno prilagaja klasifikatorje v naslednjih fazah učenja v prid vzorcem, ki niso bili obravnavani v predhodnih fazah. To ga naredi manj občutljivega na pretreniranost modela, vendar bolj občutljivega na podatke s šumom in vzorce, ki močno odstopajo od povprečne distribucije, tj. generalizirane množice.

# 4 Osnovni koncept predlaganega več-modalnega evalvatorja

## 4.1 Zahteve sistema

Pristop vrednotenja kakovosti storitve z objektivnim evalvatorjem pogojuje vrsta karakterističnih atributov evalvacijskega sistema:

- mehanizem delovanja in algoritmi procesiranja signalov, tj. zmožnosti uporabljene video in avdio metrike,
- kompleksnost in zahtevana procesorska moč,
- podobnost s testnim gradivom, tj. možnost prilagoditve v času delovanja,
- interoperabilnost,
- možnost uporabe,
- upoštevanje med-modalnih odvisnosti, itd.

Našteti atributi vplivajo na sledeče parametre vrednotenja kakovosti storitve:

- stopnja uspešnosti detekcije polj ROI,
- natančnost ocene kakovosti storitve za posamezno modalnost,
- zaznava konteksta in s tem povezana pre-nastavitev parametrov,
- delovanje v (skoraj) realnem času in
- upoštevanje med-modalnih učinkov.

Izbira ter precizna nastavitev teh atributov se nato izražata v ujemanju dobljene ocene z MOS oceno. V naši raziskavi smo uporabili orodja za procesiranje slikovnega gradiva, ki temeljijo na odprtokodni knjižnici OpenCV [11], ki smo jim dodali glasovno kakovostno metriko PESQ-WB [12]. OpenCV omogoča enostavno upravljanje, urejanje in dostopanje do video toka priključene kamere ali shranjene video datoteke. Integriran kodirnik in dekodirnik za MPEG-4/AVC in H.264 daje možnost analize in manipulacije na osnovi video struktur tj. video, slika, okvir, pol-okvir in piksel. Z izdelanimi knjižnicami imamo dostop do višje-nivojskih funkcij, tj. detekcije obraza in ust. Vključen detektor zaznava obraz na video slikah, s čimer določimo polja ROI z regijsko lokalizacijo in sledenjem, pri čemer smo uporabili obstoječ klasifikator za frontalno obrazno detekcijo. V eksperimentalnem delu smo uporabili že naučene klasifikatorje [13].

## 4.2 Testno gradivo in parametri sistema

Testno gradivo so bili posnetki zajeti iz komercialnih TV kanalov (IPTV) in spletnih HD vsebin. Posnetki imajo sledeče lastnosti:

- Video specifikacije:
  - Resolucija: HD (1920x1080i),
  - Kodek: H.264/MPEG-4 AVC,
  - Format dekodiranja: 4:2:0 YUV,
  - Osveževanje: >= 50 pol-slik/sek.,
  - Pogled: statična kamera,
  - Kontekst: reportaža ali novice (1 frontalni obraz).
- Avdio specifikacije:
  - Frekvenca vzorčenja: >= 16 kHz,
  - Resolucija: 16 bitna,
  - Število kanalov: 2,
  - Kodek: MPEG-4 AAC LC,
  - Kontekst: pretežno govor (1 govorec, nizka stopnja avdio šuma).

# 5 Performančna analiza

Pri izbiri uporabljene kakovostne metrike smo se osredotočili na kontekst testnega materiala, ki smo ga uporabili. Proces detekcije struktur obraza in ust, za katerega smo predvidevali, da bo najzahtevnejša naloga, je že v fazi preliminarnega testiranja dobro funkcioniral. Prihajalo je le do težav zaradi podvajanja detektiranih objektov , tj. več kot 1 polje ROI s strukturo »usta«, na mestu, kjer teh struktur ni bilo. Težavo smo rešili s predpostavko, da struktura »usta« obstaja le znotraj strukture »obraz«. V primeru pa, da jih detektor najde več, kot pravilno izberemo tisto, ki je bližje virtualnim koordinatam idealnega centra ust relativno znotraj strukture »obraz«, kot je podano v enačbi (4). S temi popravki smo občutno izboljšali delovanje detektorja.

$$idealniCenterUst(x,y) = \left[\left(\frac{sirinaSlikeObrazROI}{2}\right),\left(\frac{visinaSlikeObrazROI}{1.304}\right)\right] \quad (4)$$

Zaradi velike količine procesiranih podatkov (dekodiranje AV okvirja, uporaba detektorja in evalvacija SSIM (na ROI ali celem video-okvirju) ter izračun PESQ-MOS ocene) program ni deloval v realnem času, temveč počasneje (2-3 slik/sek.). Obdelavi v realnem času smo se približali (20 slik/sek.) z uporabo SSIM algoritma samo v polju ROI ter s spremenjeno nastavitvijo detektorja, kjer smo predpostavili, da imata strukturi »obraz« in »usta« določeno minimalno in maksimalno velikost. S tem smo dosegli manjše število obhodov detektorja po celotnem okvirju. Dodatno smo med preliminarnimi meritvami preizkusili še scenarij, kjer smo za video evalvacijo uporabili PSNR, vendar le do določenega praga (=40dB). Kadar je vrednost padla pod prag, smo

uporabili SSIM. Hitrost obdelave podatkov za ta pristop je variirala in je bila odvisna od stopnje degradacije testnega posnetka. Manjša kot je bila degradacija, hitreje je program deloval, pri tem pa je potrebno v zakup vzeti slabšo korelacijo PSNR z MOS.

Zvočne segmente in evalvacijo ocene PESQ-WB smo ohranili v originalni izvedbi.


Slika 1. Detekcija polj ROI "obraz" in "usta"

## 6 Zaključek

V prispevku smo predstavili model več-modalnega evalvatorja kakovosti storitve za AV visoke ločljivosti. Naredili smo preliminarne meritve in preizkus delovanja, kar bo služilo za preslikavo vrednosti in primerjavo s subjektivnimi testi. Nadaljnje delo bo potekalo predvsem v smeri določanja pomembnosti območij ROI v primeru degradiranih posnetkov zaradi vpliva prenosnega omrežja (mesto in količina degradacije v multimedijskem toku), uporabe drugih metod za detekcijo območij ROI in določanja križno-modalnih vplivov (asinhronizacija, pomembnost kanalov modalnosti).

## Zahvala

Posebna zahvala gre Evropski uniji, ki iz Evropskega socialnega sklada delno financira program usposabljanja mladega raziskovalca v okviru Operativnega programa razvoja človeških virov za obdobje 2007 – 2013.

## Literatura

[1] John G. Beerends, Frank E. de Caluwe, »The Influence of Video Quality on Perceived Audio Quality and Vice Versa«, *Audio Engineering Society*, 47(5): 355-362, 1999.

[2] O. Le Meur, A. Ninassi, P. Le Collet, »Over visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric«, *Signal Processing- Image Communication*, 25(7): 547-558, 2010.

[3] K. H. Kamrul Hasan Talukder, »Haar Wavelet Based Approach for Image Compression and Quality Assessment of Compressed Image«, *International Journal of Applied Mathematics*, 36(1), 2007.

[4] D. Venkata Rao, L. Pratap Reddy, »Image quality assessment based on perceptual structural similarity«, *Pattern recognition and machine intelligence*, Kolkata, Indija, 2007.

[5] Z. Wang et al., »Image quality assessment: From error visibility to structural similarity«, *IEEE Trans. On Image Processing*, 13(4):600-612, 2004.

[6] ITU-T, »BS.1387 : Method for objective measurements of perceived audio quality«, *http://www.itu.int/rec/R-REC-BS.1387/en*.

[7] ITU-T, »P.862: Perceptual evaluation of speech quality (PESQ)«, *http://www.itu.int/rec/T-REC-P.862*.

[8] Benjamin Belmudez, Sebastian Moeller, Blazej Lewcio, Alexander Raake, Amir Mehmood, »Audio and Visual Channel Impact on Perceived Audio-visual Quality in Different Interactive Contexts«, *Multimedia Signal Processing*, Kyoto, Japonska, 2009.

[9] D.S. Hands, »A basic multimedia quality model«, *IEEE Transactions on Multimedia*, 6(6):806-816, 2004.

[10] Yoav Freund, Robert E. Schapire, »A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting«, *Journal of computer and system sciences*, 55:119-139, 1997.

[11] Intel, »OpenCV - Open Source Computer Vision«, *http://code.opencv.org*.

[12] ITU-T, »P.862.2 : Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs«, *http://www.itu.int/rec/T-REC-P.862.2-200711-I/en*.

[13] J. M. Rainer Lienhart, »An Extended Set of Haar-like Features for Rapid Object Detection«, *IEEE International Conference on Image Processing*, Rochester, 2002.