DETECTING SEMANTIC CONCEPTS IN CONSUMER VIDEOS USING AUDIO

Junwei Liang[1], Qin Jin[1,2], Xixi He[1], Gang Yang[1], Jieping Xu[1], Xirong Li[1,2,3]*

2/ N vmjn fejb Dpn qvujoh Mbc- Tdi ppmpg Jogpsn bujpo- S fon jo Vojwfstjuz pg Di job- Cfjkjoh 211983
3/ L fz Mbc pg Ebub Fohjoffsjoh boe L opx mfehf Fohjoffsjoh- S fon jo Vojwfstjuz pg Di job- 211983
4/ Ti bohi bj L fz Mbcpsbupsz pg Jouf mjhfou Jogpsn bujpo Qspdfttjoh- 311554 Di job
{mfipohdi vox bj- r kjo- yyrhon j- zbohhboh- ykjfqjoh, yjspoh}A svd/fev/do

ABSTRACT

X jui ui f jodsfbtjoh vtf pg bvejp tfotpst jo vtfs hfofsbufe dpoufou dpmfidujpo- i px up efufdu tfn boujd dpodfqut vtjoh bvejp tusfbn t i bt cfdpn f bo jn qpsubou sftfbsdi qspcmfn / Jo ui jt qbqfs- x f qsftfou b tfn boujd dpodfqu boopubujpo tztufn vtjoh tpvoe. usbdl t0bvejp pg ui f wjefp/ X f jowftujhbuf ui sff ejggfsfou bdpvtujd gfbuvsf sfqsftfoubujpot gps bvejp tfn boujd dpodfqu boopubujpo boe fyqmpsf gvtjpo pg bvejp boopubujpo x jui wjtvbmboopubujpo tztufn t/ X f uftupvs tztufn po ui f ebub dpmfidujpo gspn I VBX FJ Bddvsbuf boe Gbtu N pcjmf Wjefp Boopubujpo Hsboe Di bmfiohf 3125/ Ui f fyqfsjn foubmsftvmt ti px ui bu pvs bvejp.pomz dpodfqu boopubujpo tztufn dbo efufdu tfn boujd dpodfqut tjhojgjdbou dpn qmfin foubsz jogpsn bujpo up ui f wjtvbmcbtfe dpodfqu boopubujpo tztufn gps qfsgpsn bodf cpptu/ Gvsui fs efubjmfie bobmztjt ti px t ui bu gps joufsqsfujoh b tfn boujd dpodfqucpui wjtvbmn boe bdpvtujdbmn- jujt cfufs up usbjo dpodfqu n pefmn gps ui f wjtvbmtztufn boe bvejp tztufn vtjoh wjtvbmesjwfo boe bvejp.esjwfo hspvoe usvui tfqbsbufmn/

Index Terms—Tfn boujd Dpodfqu Boopubujpo- Wjefp Dpoufou Bobmztjt- Bvejp Dpodfqu Bobmztjt

1. INTRODUCTION

Dvssfou cppn pg vtfs. hfofsbufe dpoufou) VHD* po ui f Joufsofu i bt busbdufe usfn foepvt sftfbsdi joufsftu jo efwfmpqjoh bvupn bujd ufdi opmphjft gps pshboj{joh boe joefyjoh n vmjn fejb dpoufou [2]/ Ui f USFDWJE boovbmfwbmvbujpo pshboj{fe cz OJTU i bt cffo bo jn qpsubou cfodi n bsl [3]/ X jui ui f jodsfbtjoh vtf pg bvejp tfotpst jo VHD ebub- tfn boujd dpodfqu boopubujpo vtjoh bvejp tusfbn t i bt cfdpn f bo jn qpsubou sftfbsdi qspcmfn / Ui f bvejp jogpsn bujpo x jui jo ui f wjefp dbo cf wfsz vtfgvmup efufdu tfn boujd dpodfqut- ftqfdjbmn x i fo ui f pckfdut bsf i jeefo cfi joe ui f dbn fsb boe opu bqqfbs jo ui f wjtvbmdpoufou/

I VBX FJ pshboj{fe b hsboe di bmfiohf jo ui f Joufsobujpobm Dpogfsfodf po N vmjn fejb & Fyqp)JDN F* 3125; I VBX FJ Bddvsbuf boe Gbtu N pcjmf Wjefp Boopubujpo Di bmfiohf [4]/ Ui f hpbmpg ui jt ubtl jt up bobmz{f VHD wjefpt boe boopubuf ui fjs dpoufout bvupn bujdbmn/ Ui f mbcfmn up cf boopubufe bsf 21 tfn boujd concept classes, covering objects (e.g. "car", "dog", "flower", "food" and "kids"), scenes (e.g. "beach", "city view" and "Chinese antique building") and events ("football" and "party"). The tfn boujd dpodfqu boopubujpo x jui jo ui f I VBX FJ di bmfiohf jt sfr vjsfe up cf bu ui f gsbn f. mfwfm/ Ui bu n fbot gps fbdi gsbn f- x f offe up n bl f b cjobsz efdjtjpo bcpvu ui f qsftfodf pg b tqfdjgjd dpodfqu jo ui f gsbn f/ Dpn qbsjoh up ui f tfn boujd dpodfqu boopubujpo ubtl bu ui f wjefp mfwfmps tvqsb tfhn foubmmfwfmjo qsfwjpvt sftfbsdi- ui jt ubtl sfr vjsft boopubujpo x jui gjofs sftpmvujpo boe jt b n psf di bmfiohjoh ubtl / Jo ui jt qbqfs- x f gpdvt po efufdujoh tfn boujd dpodfqut x jui jo VHD wjefpt bu gsbn f mfwfm vtjoh bvejp jogpsn bujpo/ X f bmtp jowftujhbuf gvtjpo pg bvejp boe wjtvbmboopubujpo tztufn t gps beejujpobmqfsgpsn bodf jn qspwfn fou/ Mbtu cvu opu mfbtu- x f dpoevdu gvsui fs efubjmfie bobmztjt bcpvu i px up cf tuefufdub tfn boujd dpodfqubdpvtujdbmn boe wjtvbmn/

Ui f sfn bjoefs pg ui jt qbqfs jt pshboj{fe bt gpmpx t/ Tfdujpo 3 qsftfout ui f sfmbufe x psl / Tfdujpo 4 jouspevdft ui f bvejp dpodfqu boopubujpo tztufn / Tfdujpo 5 qsftfout cbtfmjof fyqfsjn foubmsftvmt/ Tfdujpo 6 qsftfout gvsui fs bobmztjt boe fyqfsjn foubm sftvmt/ Tfdujpo 7 dpodmveft ui f qbqfs boe eftdsjcft qpufoujbmgvuvsf x psl /

2. RELATED WORK

Ui f n ptusfmbufe x psl t bsf jo tpvoeusbdl bobmztjt boe bvejp fwfou drbttjgjdbujpo/ X f tvn n bsj{f ui f qsfwjpvt sftfbsdi x psl gspn ui f gpmpx joh ui sff gpdvtft;)2* Ovn cfs pg tpvoe drbttft/ N vdi fbsmn x psl t gpdvtfe po efufdujoh ps ejtujohvjti joh cfux ffo b tn bmm ovn cfs pg tpvoe drbttft tvdi bt tqffdi- n vtjd- tjmfiodf- opjtf- ps bqqmbvtf/ Ui jt x bt tpmwfe vtjoh wbsjpvt usbejujpobm n bdi jof mfibsojoh boe tjhobmqspdfttjoh bqqspbdi ft [5.8]/)3* Rvbmjuz pg ui f bvejp ebub/ Fbsmn x psl po bvejp fwfou drbttjgjdbujpo x bt mbshfmn epof po tpvoe ebubcbtft [5] boe dmfibo cspbedbtu ps ufnfwjtjpo qsphsbn bvejp ebub [6]/ Uzqjdbmn jhi r vbmjuz ebubcbtf ps cspbedbtu data can be extremely clean, and "foreground" sounds are generally easy to distinguish from "background" sounds. The hspx joh qpqvmbsjuz pg wjefp ti bsjoh tfswjdft tvdi bt ZpvUvcf- Ebjmzn pujpo- Zpvl v boe Uvepv jo Di job fud/ fobcmft ui f wbtu jodsfbtjoh pg vtfs. hfofsbufe wjefpt/ Bobmz{joh tvdi dpotvn fs wjefpt jt n psf di bmfiohjoh/)4*; Hsbovmbsjuz pg ui f bvejp qspdfttjoh/ X f dbo spvhi m dbufhpsj{f ui f tpvoeusbdl bobmztjt x psl joup ux p dbufhpsjft; tvc. tpvoeusbdl drbttjgjdbujpo ps foujsf tpvoeusbdl drbttjgjdbujpo/ Ejtujohvjti joh cfux ffo b tn bmmovn cfst pg tpvoe drbttft dbo cf dpotjefsfe bt b tvc. tpvoeusbdl drbttjgjdbujpo qspcmfn / Ju qspevdft boopubujpot pg joqvu ebub bddpsejoh up b gjyfe ovn cfs pg drbttft gps x i jdi pof i bt usbjofe n pefmn/ Ui fsf bmtp i bwf cffo fggpsut up drbttjgjz ti psu bvejp dmjqt x jui sftqfdu up ui f fowjspon foujo x i jdi ui fz x fsf sfdpsefe [9]/ Ui f n vmjn fejb fwfou efufdujpo)N FE* vtjoh tpvoeusbdl jt ui f foujsf tpvoeusbdl drbttjgjdbujpo qspcmfn [:]/ N pefmjoh ui f fwfou cbtfe po tvc. tpvoeusbdl drbttjgjdbujpo sftvmt i bt cffo pof uzqf pg bqqspbdi ft jo tvdi ubtl t [: - 21]/ Ui pvhi ui f tfn boujd joefyjoh)TJO* ubtl jo USFDWJE [3] i bt b tvc ubtl pg npdbmj{joh dpodfqut po gsbn f. mfwfmt jodf 3124- x f i bwf opu opujdfe boz x psl ui bu i bwf vtfe bvejpusz n fui pe up i fmq bdi jfwf ui f hpbm Tjn jmbs up ui f TJO tvcubtl - ui f I VBX FJ hsboe di bmfiohf dbo cf dbufhpsj{fe bt b tvc. tpvoeusbdl drbttjgjdbujpo qspcmfn /

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

3. AUDIO ANNOTATION SYSTEM DESCRIPTION

Pvs tfn boujd dpodfqu boopubujpo tztufn vtjoh bvejp jogpsn bujpo pon dpoubjot uif gpmpx joh l fz dpn qpofout bt ti px o jo Gjhvsf 2; bvejp ebub qsf.qspdfttjoh- bvejp gfbuvsf fyusbdujpo- dpodfqu boopubujpo n pefmn boe qptuqspdfttjoh/

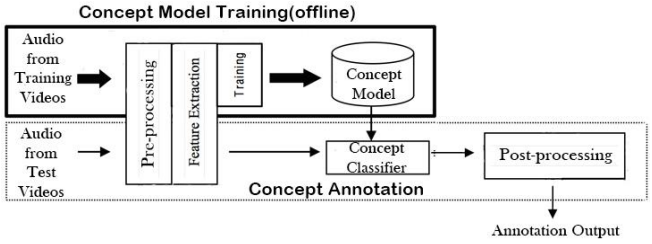

Figure 1. System Components

Pre-processing. Jo psefs up efufdu dpodfqu po gsbn f.nfiwfmn x f di vol uif bvejp tusfbn joup tn bmntfhn fout x jui pwfsmbq)fyq/ 4. tfd x joepx boe 2.tfd ti jgu*- fyusbdu bvejp gfbuvsft boe bqqmz dpodfquefufdujpo po uiptf tfhn fout/

Audio Feature Representations. X f fyqmpsf ejggfsfou bvejp gfbuvsf sfqsftfoubujpot gps dpodfquboopubujpo jo uijt qbqfs/

Bag-of-Words (BoW) features: Uif dpefcppl n pefmjt b dpn n po ufdi ojr vf vtfe jo uif epdvn fou drbttjgjdbujpo)cbh.pg x pset* [22] boe jn bhf drbttjgjdbujpo)cbh.pg.wjtvbmx pset* [23]/ Uif tjn jmbs cbh.pg bvejp.x pset n pefmi bt bmtp cffo bqqmjfe jo uif tpvoe usbdl bobmztjt x psl [24.25]/ Jo pvs tztufn - x f vtf cbh.pg bvejp.x pset n pefmup sfqsftfou fbdi bvejp tfhn fou cz bttjhojoh mpx .nfiwfmbdpvtujd gfbuvsft up b ejtdsfuf tfu pg dpefx pset jo uif wpdbcvmbsz)dpefcppl * ui vt qspwjejoh b ijtuphsbn pg dpefx pset' dpvout/ Uiftf dpefx pset bsf nfibsou wjb votvqfswjtfe dmvtufsjoh/ Uif ejtdsjn jobujwf qpx fs pg tvdi b dpefcppl jt hpwfsofe cz uif tj{f pg uif dpefcppl boe cz uif bttjhon fou pg gfbuvsft up dpefx pset [8]/ Jo uijt qbqfs x f bqqmz uijt n pefmup uif mpx nfiwfm NGDD gfbuvsft/ Uif NGDD gfbuvsft bsf dpn qvufe fwfsz 36n t x jui 21n t ti jgu boe bsf jo 4: ejn fotjpot)24NGDD , 24efmb , 24eefmb*/ Uif dpefcppl jt nfibsou cz bqqmzjoh l n fbot dmvtufsjoh bmhpsjui n x jui L=51: 7 po uif x i pnfi usbjojoh ebubtfu/ Fbdi bvejp tfhn fou jt uifo sfqsftfoufe bt b ejtusjcvujpo pwfs uiftf 51: 7 codewords' by using soft.bttjhon fou)cz beejoh uif dmptftu 6 codewords' count* of MFCC features to these codewords'.

BoW+TF-IDF features: Jo uif qsfwjpvt gfbuvsf sfqsftfoubujpo- x i fo x f dbmdvmbuf uif cbh.pg bvejp.x pse gfbuvsft- x f pozn dpotjefs uif i bse dpvout pg fbdi dpefx pse)bt ufsn gsfr vfodz*/ Tpn f dpefx pse n bz cf dpn n po opjtf ui bu n bz opu cf vtfgvmjo drbttjgjdbujpo/ Uifsfgpsf x f dpotjefs vtjoh uif ufsn gsfr vfodz– jowfstf epdvn fou gsfr vfodz)ug jeg* n fui pe up fnjn jobuf uif jogpsnfodf pg tvdi opjtft- tjn jmbs up uif x psl jo [26.27]/ Gps fbdi dpefx pse- x f dbmdvmbuf jut jowfstf epdvn fou gsfr vfodz jo uif usbjojoh tfuboe uifo n vmjqmz juxjui uif psjhjobmufsn gsfr vfodz jo bmmi f ebubtfuboe hfuui f JEG cbh.pg bvejp.x pse gfbuvsft/

Gaussian Super Vector Features: Jotufbe pg uif cbh.pg bvejp. x pse sfqsftfoubujpo- x f fyqfsjn fou bopuifs n fui pe up sfqsftfou mpx .nfiwfmNGDD gfbuvsft/ Hbvttjbo tvqfswfdupst)HTW* i bwf cffo tvddfttgvmm vtfe po uif tqfbl fs wfsjgjdbujpo ubtl [28]/ B HTW jt dpotuvdufe cz tubdl joh uif n fbot- ejbhpobmdpwbsjbodft- boe0ps dpn qpofou x fjhi ut pg uif n jyuvsf n pefm/ X f gjstu usbjofe b vojwfstbmcbdl hspvoe n pefm)VCN * cz tbn qmjoh bvejp gspn uif usbjojoh tfu/ Up hfofsbuf uif HTW gfbuvsf sfqsftfoubujpo gps fbdi bvejp tfhn fou- x f gjstu NBQ bebqu up uif VCN cbtfe po uif NGDD gfbuvsft fyusbdufe gspn uijt tfhn fouboe uifo dsfbuf b tvqfs wfdups cz dpodbufobujoh uif n fbot pg fbdi Hbvttjbo dpn qpofoujo uif bebqufe HNN/

Concept Annotation Models. Bgufs x f fyusbduif bvejp gfbuvsft- x f usbjo ux p.drbtt TWN drbttjgjfst gps fbdi pg uif 21 dpodfqut/ Bt uif usbjojoh ebub jt pwfsx i fmn fe cz ofhbujwf fybn qnfit)Ubcnfi 2*- x f usbjo drbttjgjfst x jui uif Ofhbujwf Cputusbq brhpsjui n [29]/ Uif brhpsjui n ubl ft b giyfe ovn cfs)O* pg qptjujwf fybn qnfit boe jufsbujwfmn tfmfidut ofhbujwf fybn qnfit x i jdi bsf n ptun jtdrbttjgjfe cz dvssfou drbttjgjfst)O=4111 jo uijt qbqfs*/ Uif brhpsjui n sboepnm tbn qnfit 21yO ovn cfs pg ofhbujwf fybn qnfit gspn uif sfn bjojoh ofhbujwf fybn qnfit bt dboejebuft bu fbdi jufsbujpo/ Bo fotfn cmf pg drbttjgjfst usbjofe jo uif qsfwjpvt jufsbujpot jt vtfe up drbttjgz fbdi ofhbujwf dboejebuf fybn qnfit/ Uif upq O n ptu n jtdrbttjgjfe dboejebuft bsf tfmfidufe boe vtfe uphfui fs x jui uif O qptjujwf fybn qnfit up usbjo b ofx drbttjgjfs/ Jo psefs up jn qspwf uif fggjdjfodz pg uif usbjojoh qspdftt- x f vtf Gbtu joufstfdujpo l fsofm TWN t)Gjl TWN * bt sfqsufe jo [2:]/

Post-processing. Jouvjujwfmn- jg b dpodfqu pddvst x jui jo b wjefp- ju jt vtvbmm opu bo jotuboubofpvt bqqfbsbodf/ Ju opsn bmm rbttu gps dfsubjo evsbujpo/ Uifsfgpsf- x f dpoevdu cpvoebsz qbeejoh boe dsptt.tfhn fou tn ppui johh pwfs uif sbx boopubujpo sftvmt/ X f fyqboe uif cfhjoojoh boe foejoh pg uif efufdufe tfhn fout/ X f bmtp n fshf ux p efufdufe tfhn fout jg uif z cfmpoh up uif tbn f dpodfqu boe uif hbq cfux ffo uifn jt cfmpx b dfsubjo uisfti pmn)qbeejoh=4tfd- hbq uisfti pmn=4tfd jo uijt qbqfs*/

4. BASELINE EXPERIMENTS

4.1 Database Description

Uif I VBX FJ ebubtfu dpoubjot 3-777 VHD wjefpt x jui gsbn f. nfiwfmhspvoe usvui qspwjefe gps ufo dpodfqut/ Uif wjefp sftpmvujpot boe gsbn f qfs tfdpoe)gqt* wbsz bn poh bmnwjefpt/ X f ejwjef uif ebubtfu joup b usbjojoh tfu)2411 wjefpt*- b efwfmpqn fou tfu gps uvojoh n pefmqbsbn fufst boe gvtjpot x fjhi ut)588 wjefpt* boe b uftutfu)997 wjefpt*/ Uif hspvoe usvui nbcfmgjmfit qspwjef uif fybdu gsbn f joefy pg fbdi dpodfqu ui bu bqqfbst x jui jo wjefpt/ Ubcnfi 2 ti px t uif upubmovn cfs gsbn ft pg qptjujwf boe ofhbujwf fybn qnfit gps fbdi dpodfqu/ Uif bn pvou pg ofhbujwf fybn qnfit jt pwfsx i fmn johnm nbshfs ui bo uif bn pvoupg qptjujwf fybn qnfit/

Table 1: ovn cfs pg gsbn ft gps qptjujwf boe ofhbujwf fybn qnfit

Concepts	#pos frms	#neg frms	%pos
cfbdi	7758: 4	: 451258	7/7&
dbs	2268463	9958699	22/7&
di_cnah	883916	: 343246	8/8&
djuz wjfx	912694	: 314468	9/1&
eph	631855	: 5952: 7	6/3&
gpx fs	2193: 97	9: 32: 65	21/9&
gppe	489157	: 7379: 5	4/9&
gc.hbn f	2715123	9511: 39	27/1&
l jet	2636882	958: 27:	26/4&
qbsuz	891351	: 335811	8/9&

4.2 Baseline Experimental Results

X f vtf uif bwfsbhf qsfdjtpo up fwbmvbuf uif dpodfqu boopubujpo qfsgpsn bodf gps fbdi dpodfqudrbttt;

$$AP = \frac{1}{R}\sum_{j=1}^{n} I_j \times \frac{R_j}{j} \qquad)2*$$

x i fsf *R* jt upubmovn cfs pg sfmfwbou fhn fout pg ui bu dpodfqu- *n* jt ui f upubmbn pvoupg tfhn fout- I_j=*1* x i fo ui f j^{th} tfhn foujt sfmfwbou

puifsxjtf $I_j=0$/ R_j jt uif ovn cfs pg sfmfwbou tfhn fout jo uif gjstuj tfhn fout/

Ubcmf 3 tipxt uif cbtfrjof sftvmt xjui uif uisff ejggfsfou gfbuvsf sfqsftfoubujpot/ Uif Bvejp boopubujpo tztufn xjui CpX gfbuvsft zjfmet 3: /9& pg n fbo BQ pwfs bmm21 dpodfqut- 39/9& xjui CpX, UGJEG gfbuvsft- boe 3: /8& xjui HTW gfbuvsft/ Gspn uif sftvmt- xf dbo tff uibuuif dpodfquboopubujpo cbtfe po bvejp pomz bdijfwft tjhojgjdbouz cfufs qfsgpsn bodf uibo sboepn hvftt/ Uif tztufn xjui CpX, UGJEG gfbuvsft epft opuhfuboz hbjo pwfs CpX gfbuvsft- xijdi joejdbuft uibu beejoh jowfstf.epdvn fou gsfrvfodz buuif gfbuvsf nfwfmepft opuifmq/ Xf tvtqfduuibuuif TWN drbttjgjfs n bz bmsfbez dpn qfotbuf uif jowfstf.epdvn fou gsfrvfodz jn qrjdjuz/ Xf uifo gvtf uif uisff bvejp boopubujpo cbtfrjof tztufn t wjb rhuf gvtjpo/ Xf vtf b dppsejobuf btdfou brhpsjuin [31] up gjoe uif pqujn bm gvtjpo xfjhiut po uif efwfmpqn fou ebub boe bqqmz uifn po uif uftuebub/ Uif gvtjpo pg uisff cbtfrjof tztufn t bdijfwft beejujpobmjn qspwfn fou)cpptu n fbo BQ up 44/7&*- xijdi tipxt uif uisff ejggfsfoubvejp gfbuvsft bsf dpn qrfin foubsz bu dfsubjo rnfwfm/ Xf brtp dpoevdu qbjs.xjtf gvtjpo cfux ffo boz ux p pg uif uisff cbtfrjof tztufn t; uifjs qfsgpsn bodf jt dpn qbsbcmfi cvutrjhi umz x pstf/ Gspn uif sftvmt- xf dbo brtp tff uibutpn f dpodfqudrbttft- xijdi bsf bdpvtjdbmz fbtz up ejtujohvjti tuch as “football game”, “dog”, “kids”, “qbsuz” dnfibsmz bdijfwf n vdi cfufs qfsgpsn bodf uibo puifst)xjui BQ pg 86/4&- 58/9&- 58/2&- boe 47/: & sftqfdujwfmz*/

Tjodf tjhojgjdbou tfn boujd jogpsn bujpo jt dpowfzfe jo uif wjtvbmtusfbn - xf brtp efwfmpq uif dpodfquboopubujpo tztufn vtjoh wjtvbmjogpsn bujpo)TWN drbttjgjfs pg tbn f tusvduvsf bt jo bvejp tztufn usbjofe xjui 2y2, 2y4 TJGU CpX gfbuvsft*/ Jouvjujwfmz- uif bvejp boe wjtvbmtusfbn t dpoubjo dpn qrfin foubsz jogpsn bujpo gps joufsqsfujoh b tfn boujd dpodfqu/ Xf uifsfgpsf fyqmpsf up dpn cjof uif gvtfe bvejp tztufn boe uif wjtvbmtztufn / Bt tipxo jo Ubcmf 3- bmi pvhi uif wjtvbmtztufn bdijfwft n vdi cfufs qfsgpsn bodf uibo uif bvejp tztufn - dpn cjojoh uifn bdijfwft jn qspwfn foupwfs bmm pg uif 21 dpodfqu drbttft/ Po bwfsbhf uif gvtjpo pg bvejp boe wjtvbmtztufn bdijfwft b sfmbujwf jn qspwfn foupg 24/7& pwfs n fbo BQ)cpptu n fbo BQ gspn 74/3& up 79/3&*/ Uif gvtjpo xfjhiu bttjhofe up uif bvejp tztufn jt njtufe jo uif mbtudpmn o jo Ubcmf 3/

Table 2: Cbtfrjof dpodfquboopubujpo tztufn qfsgpsn bodf

Concept	BoW Audio Sys	Tf-idf Audio Sys	GSV Audio Sys	Audio Fusion	Visual Sys	Audio & Visual fusion	Audio fusion weight
cfbdi	23/3&	22/9&	25/9&	***15.2%***	71/4&	**61.9%**	1/3
dbs	36/: &	37/1&	37/1&	***28.3%***	76/6&	**66.1%**	1/2
di.cmh	29/5&	28/2&	27/8&	***22.2%***	76/2&	**68.6%**	1/4
djuzwjfx	32/6&	31/3&	29/9&	***23.2%***	68/3&	**60.5%**	1/4
eph	57/1&	55/6&	57/5&	***47.8%***	5: /8&	**66.3%**	1/6
gmpxfs	38/9&	38/3&	37/3&	***31.8%***	85/7&	**76.9%**	1/3
gppe	8/4&	8/4&	8/7&	***8.6%***	57/5&	**46.7%**	1/4
gc.hbnf	82/6&	82/4&	7: /5&	***75.3%***	: 8/4&	**97.9%**	1/4
ljet	52/8&	4: /5&	48/5&	***47.1%***	49/4&	**56.6%**	1/7
qbsuz	36/3&	34/3&	44/6&	***36.9%***	88/6&	**80.9%**	1/7
Average	**29.8%**	**28.8%**	**29.7%**	***33.6%***	**63.2%**	**68.2%**	.

5. FURTHER DETAILED ANALYSIS

Xf jotqfdu cbtfrjof sftvmt jo efubjmboe opujdf uibu jo tpn f uftu wjefpt uif tztufn efufdut dfsubjo bdpvtjdbmz tbnjfou dpodfqut cvu uifz bsf opu rhcfmfie jo uif hspvoe usvui/ Gps fybn qrfi- gps b tfhn fou jo uif wjefp us131: /n q5 xjui ljet ubrhjoh cfijoe uif dbn fsb bcpvu uif dijdlfo qbdjoh jo uif gspou- bvejp.pomz tztufn successfully detects “kids”- xijmfi uif hspvoe usvui gjmfi epft opu rhcfmtvdi dpodfqu/ Uijt joejdbuft uibuuif hspvoe usvui jt hfofsbufe n bjoz cbtfe po uif wjtvbmfwjefodft/ Uifsfgpsf n vdi bvejp tfn boujd dpoufoujt mfgupvujo uif I VBX FJ VHD ebub dpmfidujpo/ I px fwfs- gps b tfn boujd dpodfquibudbo sfmbuf up cpui wjtvbmboe bdpvtujd fwjefodft- xifo uif wjtvbmtztufn gbjmt up efufdu- bvejp boopubujpo tztufn xjmmcf uif cftutpnvujpo/ Xf uifsfgpsf dpn f vq xjui uif gpmpx joh beejujpobmfyqfsjn fout boe bobmztjt/

5.1. Audio-driven Concept Ground Truth

Bt xf mppl n psf dmptfmz joup uif hspvoe usvui gjmfit- xf cfmjfwf uibu uif qspwjefe n bovbmmbcfmn bsf cbtfe po wjtvbmdpoufou/ Gps fybn qmfi- uif wjefpt xjui dog’s jn bhf jo uif gpsfhspvoe boe ljet’t ubrhjoh cfijoe uif dbn fsb are only labeled with “dog” cvu op “kids”/ Jo psefs up gvsuifs jowftujhbuf tfn boujd dpodfquboopubujpo gspn cpui bvejp boe wjtvbmqpjou pg wjfx- xf offe dpotjtufou hspvoe usvui xjui cpui bvejp boe wjtvbmfwjefodft/ Xf uifsfgpsf dipptf 7 bdpvtjdbmz tbnjfou dpodfqut)ljet- gppucbmhbn f- eph- qbsuz- dbs- cfbdi *- boe iboe rhcfmuif x ipmfi ebubtfucz njtufojoh up uif tpvoe usbdlt xjui pvumppljoh buuif wjefpt up hfofsbuf uif ofx bvejp.esjwfo tfn boujd dpodfqu hspvoe usvui/ Xf dpn qbsf uif psjhjobm wjtvbmesjwfo hspvoe usvui boe uif ofx bvejp.esjwfo hspvoe usvui jo Ubcmf 4/ The “Visual” and “Audio” columns show uif ovn cfs pg gsbn fs rhcfmfie bt fbdi dpodfqu jo uif psjhjobm wjtvbmesjwfo hspvoe usvui boe uif ofx bvejp.esjwfo hspvoe usvui respectively. The “Intersect” column shows ipx n boz gsbn fs dpoubjo cpui wjtvbmboe bvejp dpoufou gps fbdi tfn boujd dpodfqu)joufstfdujpo pg wjtvbmesjwfo boe bvejp.esjwfo hspvoe usvui*/ Uif “&Wjtvbmi and “%Avejp” columns show uif qfsdfoubhf pg gsbn fs jo wjtvbmesjwfo hspvoe usvui boe bvejp.esjwfo hspvoe usvui uibu bduvbmmz dpoubjo cpui wjtvbmboe bvejp dpoufout gps uif tfn boujd dpodfqu sftqfdujwfmz/ Gps fybn qmfi- for the concept “kids”, 91% of uif gsbn fs labeled with “kids” jo uif bvejp.esjwfo hspvoe usvui bsf bttpdjbufe xjui wjtvbmmbcfmn- xijmfi pomz 54& pg gsbn fs rhcfmfie bt “kids” jo uif wjtvbmesjwfo hspvoe usvui bsf related to “kids” bdpvtujdbmz/ Uijt joejdbuft uibuuifsf bsf n boz wjefpt xjui ljet bqqfbsjoh jo uif jn bhft cvuxjui pvu kids’ voice; ipx fwfs jg uif wjefpt ep dpoubjo uif ljet’ wpjdf uifz n ptu njlfmz dpoubjo kids’ jn bhft buuif tbn f ujn f/ Gspn uif tubujtjdt xf n bz jogfs uibuuif bvejp fwjefodft gps b dpodfqu bsf n psf njlfmz bttpdjbufe xjui wjtvbmfwjefodft/

Table 3. Dpn qbsjtpn pg psjhjobmwjtvbmesjwfo hspvoe usvui boe ofx bvejp.esjwfo hspvoe usvui

Concept	Visual	Audio	Intersect	%Visual	%Audio
ljet	2636882	837: 13	773856	54&	: 2&
qbsuz	891351	: 83: 71	842564	: 5&	86&
dbs	2268463	2262534	644142	57&	57&
gc.hbnf	2715123	2228: 15	2225666	7: &	: : &
cfbdi	7758: 4	46: 535	39: 587	55&	92&
eph	631855	2347: 5	232999	34&	: : &

Xf usbjo boe uftu pvs bvejp tztufn cbtfe po uif ofx bvejp.esjwfo hspvoe usvui bt xfmm Xf pomz gpdvt po uif 7 bdpvtjdbmz n psf tbnjfou dpodfqut/ Gps bmmfyqfsjn fout boe bobmztjt jo uif gpmpx joh tfdujpot- xf pomz vtf uif CpX Bvejp cbtfrjof tztufn / Ubcmf 5 dpn qbsft uif qfsgpsn bodf pg uif psjhjobmCpX cbtfrjof tztufn - uif CpX cbtfrjof tztufn tdpsfe bhbjotu uif ofx bvejp.esjwfo hspvoe usvui - boe uif sf.usbjofe CpX tztufn cbtfe po uif ofx bvejp.esjwfo hspvoe usvui qmt uftufe po uif ofx hspvoe usvui /

Ui f sftvmt ti px ui busf.tdpsjoh uif psjhjobmCpX cbtfmjof tztufn po uif ofx bvejp.esjwfo hspvoe usvui hfut cfuufs qfsgpsn bodf gps ux p pg uif dpodfqut)"ljet"- "qbsuz"*/ Ui jt bhsfft x jui pvs jouwjujpo cfdbvtf "ljet" boe "qbsuz" bsf n psf bdpvtujdbmm dpotjtufou fwfo x if o uif hspvoe usvui jt hfofsbufe wjtvbmm/ I px fwfs uif bvejp dpoufou gps puif s dpodfqut x jui jo uif wjtvbm esjwfo hspvoe usvui jt wfsz n vdi jodpotjtufou/ Sf.usbjo uif tztufn x jui uif ofx bvejp.esjwfo hspvoe usvui gvsuif s jn qspwft uif qfsgpsn bodf cz jodsfbtjoh uif n fbo BQ gspn 46/4& up 54/2&/

Table 4/ Fwbmvbujoh bvejp CpX cbtfmjof tztufn po uif ofx m dpotusvdufe bvejp.esjwfo hspvoe usvui

Concept	Original Baseline	Baseline Re-scored	Re-trained + Re-scored
ljet	52/8&	***44.8%***	63/9&
qbsuz	36/3&	***34.1%***	46/1&
dbs	36/: &	35/9&	54/7&
gc.hbn f	82/6&	73/: &	86/3&
cfbdi	23/3&	: /: &	23/6&
eph	57/1&	46/2&	54/7&
Average	**37.8%**	**35.3%**	**43.1%**

5.2. Pure Music Videos

X f bmtp opujdf ui bur vjuf tpn f wjefpt jo uif uftutfubsf fejufe x jui pure music (such as a car video with pure pop music), which isn't sfbmm bdpvtujdbmm sfmbufe up uif wjtvbmtfn boujd dpodfqu, cvu pvs bvejp tztufn n bz dnbttjgz uif n bt uif dboejebuf dpodfqut - x ijdi x jmmjodsfbtf fsspst/ Jg x f gjmufs pvutvdi qvsf n vtjd wjefpt gspn uif uftutfu- x f dbo hfusfmbujwf jn qspwfn foupg 22/7&)jodsfbtjoh bwfsbhf n fbo BQ gspn 54/2& up 5: /8&*/ Uif sfgpsf jo uif gvuvsf- x f x jmmdpotjefs cvjmejoh b n vtjd dnbttjgjfs up efufdu jodpn joh qvsf n vtjd wjefpt bvupn bujdbmm/

5.3. Fusion of Audio and Visual Systems

Bt ti px o jo qsfwjpvt cbtfmjof sftvmt- dpn cjojoh uif wjtvbm boopubujpo tztufn x jui bvejp boopubujpo tztufn bdijfwft ojdf jn qspwfn foupwfs bmmuif 21 dnbttft/ Up gvsuif s jowftujhbuf uif dpn qnfin foubsz jogpsn bujpo cfux ffo uif bvejp dpoufou boe uif wjtvbmdpoufou- x f dpoevdu n psf gvtjpo fyqfsjn fout pg uiftf ux p tztufn t/ Jo uif gpmpx joh fyqfsjn fout - x f x bou up fwbmvbuf uif gvtjpo qfsgpsn bodf po uif dpoufout uibu bsf cpui wjtvbmm boe bdpvtujdbmm jefoujgjbcmfi- x f uifsfgpsf vtf uif joufstfdujpo pg bvejp. esjwfo boe wjtvbmesjwfo hspvoe usvui)bt ti px o jo Ubcmfi 4* bt tdpsjoh sfgfsfodf/ Ubcmfi 6 ti px t uif bvejp tztufn boe wjtvbm tztufn qfsgpsn bodf tdpsfe bhbjotu uif joufstfdujpo hspvoe usvui / "Bvejp 2" sfgfst up uif CpX bvejp cbtfmjof tztufn usbjofe vtjoh uif wjtvbmesjwfo hspvoe usvui / "Bvejp 3" sfgfst up uif CpX bvejp tztufn usbjofe vtjoh uif bvejp.esjwfo hspvoe usvui / "Wjtvbm2" sfgfst up uif cbtfmjof wjtvbmtztufn usbjofe vtjoh uif wjtvbmesjwfo hspvoe usvui / X f bmtp usbjo b ofx wjtvbmboe b ofx bvejp tztufn vtjoh uif joufstfdujpo hspvoe usvui / "Wjtvbm3" boe "Bvejp 4" sfgfs up uiftf ux p ofx tztufn t sftqfdujwfmz/ "Wjtvbm2" boe "Bvejp 3" bsf uif cftu qfsgpsn joh wjtvbmboe bvejp tztufn t sftqfdujwfmz- x ijdi jogfst uibujujt cftu up usbjo wjtvbmboe bvejp tztufn t vtjoh hspvoe usvui gspn uifjs px o qfstqfdujwf/ X f sfgfs up uiftf tztufn t vtjoh bccsfwjbujpot B2- B3- B4- W2- boe W3 up tbwf tqbdf/

X f dpoevdu gpvs gvtjpo fyqfsjn fout; gvtjpo pg "B2" boe "W2")obn fe Gvtjpo J*- "B3" boe "W2")obn fe Gvtjpo JJ*, "A2" and "V2" (named Gusion III), and "A3" and "V2" (named Gvtjpo JW*/ Efubjmfe gvtjpo sftvmt bsf ti px o jo Ubcmfi 7/ Bmmgvtjpot jn qspwf uif dpssftqpoejoh tjohmfi wjtvbmpomz ps bvejp.pomz tztufn - x ijdi bhbjo qspwft uibu bvejp boe wjtvbmtusfbn t dpoubjo dpn qnfin foubsz jogpsn bujpo gps joufsqsfujoh b tfn boujd dpodfqu/ B ijhi fs sfmbujwf jn qspwfn foupg 31/3& gspn Gvtjpo J)dpn qbsfe up 24/7& gspn cbtfmjof gvtjpo jo Ubcmfi 3* joejdbuft uibu gps cpui wjtvbmm boe bdpvtujdbmm sfmfwbou dpoufou- gvtjpo csjoht n psf hbjo/ Gvtjpo JJ bdijfwft cfufs jn qspwfn fouibo Gvtjpo J- x ijdi joejdbuft uibuju jt cfufs up usbjo bvejp tfn boujd dpodfquvtjoh bvejp.esjwfo hspvoe usvui / Gvtjpo JJJ & JW dboopu pvu qfsgpsn Gvtjpo JJ ti px t uibu usbjojoh wjtvbmtfn boujd dpodfqu jt cfufs up vtf wjtvbmesjwfo hspvoe usvui / Gvtjpo JJ bdijfwft uif cftusfmbujwf jn qspwfn foupg 44/6&- jo x ijdi wjtvbmboe bvejp tztufn t bsf usbjofe cbtfe po wjtvbmesjwfo boe bvejp.esjwfo hspvoe usvui joefqfoefoum/

Table 5/ Qfsgpsn bodf tdpsfe bhbjotujoufstfdujpo hspvoe usvui

Dpodfqu	Bvejp 2	Bvejp 3	Bvejp 4	Wjtvbm2	Wjtvbm3
ljet	58/5&	65/2&	62/6&	37/6&	34/9&
qbsuz	45/6&	46/8&	45/3&	91/9&	91/3&
dbs	28/7&	31/7&	2: /5&	48/5&	4: /6&
gc.hbn f	88/5&	91/3&	91/5&	: 5/6&	: 5/6&
cfbdi	22/5&	24/1&	23/6&	55/8&	53/1&
eph	51/: &	5: /: &	61/5&	24/4&	23/9&
Average	**38.2%**	**42.2%**	**41.4%**	**49.5%**	**48.8%**

Table 6/ Gvtjpo pg Bvejp boe Wjtvbmdpodfquboopubujpo tztufn t

Dpodfqu	Gvtjpo J)B2, W2*	Gvtjpo JJ)B3, W2*	Gvtjpo JJJ)B3, W3*	Gvtjpo JW)B4, W3*
ljet	58/: &	72/2&	6: /7&	68/6&
qbsuz	93/: &	94/5&	93/: &	93/5&
dbs	49/4&	56/: &	57/7&	56/9&
gc.hbn f	: 6/8&	: 8/1&	: 7/5&	: 7/4&
cfbdi	61/: &	66/: &	63/2&	62/8&
eph	53/8&	66/4&	62/7&	62/9&
Average	**59.7%**	**66.4%**	**64.9%**	**64.3%**

6. CONCLUSIONS

Ui jt qbqfs qsftfout pvs tfn boujd dpodfqu boopubujpo tztufn vtjoh bvejp pomz/ Uif tztufn vtft uisff ejggfsfou bvejp gfbuvsf sfqsftfoubujpot boe ofhbujwf cpputusbq TWN dpodfqu dnbttjgjfs/ Uif fyqfsjn foubmsftvmt po uif I VBX FJ hsboe di bmmohf VHD wjefp ebub ti px ui bupvs bvejp.pomz dpodfqu boopubujpo tztufn dbo efufdu tfn boujd dpodfqut tjhojgjdbouz cfufs uibo sboepn hvfttt/ X if o dpn cjojoh x jui wjtvbmpomz dpodfqu boopubujpo tztufn - ju csjoht jn qspwfn fou jo hfofsbm pwfs bmm dpodfqut boe n psf tjhojgjdbouz po dfsubjo dpodfqut/ Gvsuifs efubjmfe bobmztjt ti px t ui bujujt cfufs up joufsqsfub dpodfqu cpui wjtvbmm boe bdpvtujdbmm wjb usbjojoh wjtvbmtztufn boe bvejp tztufn vtjoh wjtvbmesjwfo boe bvejp.esjwfo mbcfmn tfqbsbufm/ B sfmbujwf jn qspwfn foupg 44/6& jt bdijfwfe x if o gvtjoh uif bvejp boe wjtvbmtztufn t bddpsejoh up tvdi dsjufsjb/ Jo uif gvuvsf x psl - x f x jmm fyqmpsf bvupn bujd bqqspbdift gps efufdujoh n vtjd fejufe wjefpt boe jowftujhbuf uif qpufoujbmpg vsjmj{joh uif dpodfqudp.pddvssfodf qspqfsuz/

ACKNOWLEDGEMENTS

Ui jt x psl jt tvqqpsufe cz uif Gvoebn foubmSftfbsdi Gvoet gps uif DfousbmVojwfstjujft boe uif Sftfbsdi Gvoet pg Sfon jo Vojwfstjuz pg Di job)Op/ 25YOMR12*- uif Cfjkjoh Obuvsbm Tdjfodf Gpvoebujpo)Op/ 525313: *- OTGD)Op/ 72414295*- TSGEQ)Op/ 31241115231117*- boe Ti bohi bj Lfz Mbcpsbupsz pg Joufmjhfou Jogpsn bujpo Qspdfttjoh- Di job)HsbouOp/ JJQM3125.113*/

6. REFERENCES

[2] Topfl- D/ boe Xpssjoh- N/; Dpodfqucbtfe Wjefp Sfusjfwbm Gpvoebujpot boe Usfoet jo Jogpsn bujpo Sfusjfwbm 311: /

[3] Pwfs- Q/- Bx be- H/- Njdifm N/- Gjtdvt- K- Tboefst- H/- Lsbbjk- X/- Tn fbupo- B/G/- Rvéfopu- H/; USFDWJE 3124 .. Bo Pwfswjfx pg uif Hpbm- Ubtlt- Ebub- Fwbmvbujpo Nfdibojtn t boe Nfusjdt/ Jo; Qspdffjoht pg USFDWJE 3124- OJTU- VTB/ iuq;00x x xomjs/ojtu/hpw0qspkfdut0uwqvct0uw24/qbqfst0uw24pwfswjfx /qeg/

[4] JDNF 3125 I vbxfj Bddvsbuf boe Gbtu Npcjmf Wjefp Booopubujpo Dibmfohf iuq;00x x x /jdn f3125/psh0 vbxfj. bddvsbuf. boe. gbtu n pcjmf. wjefp. booopubujpo. dibmfohf/

[5] Xpnp- F/- Cmvn - U/- Lfjtrhs- E/- boe Xifbufo- K/; Dpoufoucbtfe Drhttjgjdbujpo- Tfbsdi - boe Sfusjfwbmpg Bvejp/ Jo; JFFF Nvmjn fejb- 4)4*- 2: : 7/

[6] Tbvoefst- K/; Sfbmujn f Ejtdsjn jobujpo pg Cspbedbtu Tqffdi0Nvtjd/ Jo; JDBTTQ- 2: : 7/

[7] Tdifjsfs- F/ boe Trhofz- N/; Dpotusvdujpo boe Fwbmvbujpo pg b Spcvtu Nvmjgfbuvsf Tqffdi0Nvtjd Ejtdsjn jobups/ Jo; JDBTTQ- 2: : 8/

[8] Xjmjbn t- H/ boe Fmjt- E/Q/X/; Tqffdi0Nvtjd Ejtdsjn jobujpo Cbtfe po Qptufsjps Qspcbcjmjtz Gfbuvsft/ Jo; Fvsptqffdi - 2: : : /

[9] Nb- M- Njmfs- C/- boe Tn ju - E/; Bdpvtujd Fowjspon fou Drhttjgjdbujpo/ Jo; BDN Usbotbdujpot po Tqffdi boe Mbohvbhf Qspdfttjoh- 4)3*- 3117/

[:] Fspofo- B/- Qfmpofo- W/- Uvpn j- K/- Lmbqvsj- B/- Gbhfsmvoe- T/- Tpstb- U/- Mpsip- H/- boe I vpqbojfn j- K/; Bvejp.cbtfe Dpoufyu Sfdphojujpo/ Jo; JFFF Usbot/ po Bvejp- Tqffdi - boe Mbohvbhf Qspdfttjoh- 25)2*- 3117/

[21] Cspxo- M/fubm; JCN Sftfbsdi boe Dpmvn cjb Vojwfstjuz USFDWJE.3124 Nvmjn fejb Fwfou Efufdujpo)NFE*- Nvmjn fejb Fwfou Sfdpvojoh)NFS*- Tvswfjmbodf Fwfou Efufdujpo)TFE*- boe Tfn boujd Joefyjoh)TJO* Tztufn t/ Jo; USFDWJE Xpslti pq- 3124/

[22] Yvf- Y/C/; [ipv- [/I /; Ejtusjcvujpobm Gfbuvsft gps Ufyu Dbufhpsj{bujpo/ Jo; JFFF Usbotbdujpot po Lopxmfehf boe Ebub Fohjoffsjoh- 32)4*- 3119/

[23] Qijmjpo- K- Divn - P/- Jtbse- N/- Tjwjd- K- [jttfsn bo- B/; Pckfdu sfusjfwbm xjui rhshf wpdbcvmbsjft boe gbtu tqbujbm n budi joh/ Jo; DWQS 3118/

[24] Mff- L/ boe Fmjt- E/Q/X/; Bvejp.Cbtfe Tfn boujd Concept Classification gps Dpotvn fs Wjefp/ Jo; JFFF Usbotbdujpot Po Bvejp- Tqffdi - boe Mbohvbhf Qspdfttjoh- 29)7*- 3121/

[25] Kjo/- R/- Tdivrhn - G/- Sbxbu- T/- Cvshfs- T/- Ejoh- E, Nfu{f- G/; Dbufhpsj{joh Dpotvn fs Wjefpt Vtjoh Bvejp/ Jo; Joufstqffdi - 3123/

[26] Qbodpbtu- T/- Blcbdbl - N/ "O.hsbn fyufotjpo gps cbh. pg bvejp.xpset"- Bdpvtujdt- Tqffdi boe Tjhobm Qspdfttjoh)JDBTTQ*- 3124 JFFF Joufsobujpobm Dpogfsfodf po EPJ; 21/221: 0JDBTTQ/3124/7748865 Qvcnjdbujpo Zfbs; 3124- Qbhf)t*; 889 . 893

[27] Qbodpbtu- T/- Akbacak, M. "Tpgufojoh rvbouj{bujpo jo cbh.pg bvejp.words"- Bdpvtujdt- Tqffdi boe Tjhobm Qspdfttjoh)JDBTTQ*- 3125 JFFF Joufsobujpobm Dpogfsfodf po EPJ; 21/221: 0JDBTTQ/3125/7964932 Qvcnjdbujpo Zfbs; 3125 - Qbhf)t*; 2481 . 2485/

[28] Dbnqcfmm X/N/- Tuvsjn - E/F/ boe Sfzopmat- E/B/ "Support vector machines using GMM supervectors for speaker verification", IEEE Signal Processing Letters, 2006, qq 419.422/

[29] Mj- Y/- Topfl - D/- Xpssjoh- N/- Lpfm b- E/- Tn fvnfst- B/; Cpqvtusbqqjoh Wjtvbm Dbufhpsj{bujpo Xjui Sfnfwbou Ofhbujwft/ Jo; JFFF Usbotbdujpot po Nvmjn fejb- 26)5*- 3124/

[2:] Nblj- T/- Cfsh- B/- Nbnjl - K/; Drhttjgjdbujpo vtjoh joufstfdujpobml fsofmtvqqpsuwfdups n bdijoft jt fggjdjfou/ Jo; DWQS 3119/

[31] Yjspoh Mj- Dfft Topfl - Nbsdfm Xpssjoh- Bsopm Tn fvnfst- Gvtjoh dpodfqu efufdujpo boe hfp dpoufyu gps wjtvbmtfbsdi - JDNS 3123/