Samodejno razpoznavanje govora: od unimodalnih k bimodalnim sistemom

Anton Pozne ml., Nikola Pavešić, France Mihelič

Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška 25, 1000 Ljubljana
{tonep,nikolap,mihelicf}@luks.fe.uni-lj.si

Povzetek

Predvsem v hrupnejšem okolju si ljudje pri govorni komunikaciji učinkovito pomagamo tudi z vidno informacijo o sogovornikovem obrazu. Še posebej pozorno spremljamo gibanje sogovornikovih ustnic. Tudi pri samodejnem razpoznavanju govora lahko poleg akustične informacije v sistem vodimo tudi vidno informacijo o govorčevem obrazu. V prispevku na kratko obravnavamo ključne elemente oziroma postopke, ki sestavljajo bimodalne (avdio-vizualne) sisteme za samodejno razpoznavanje govora. Poudarjen je predvsem postopek obdelave zajetih vizualnih podatkov, samodejno pridobivanje primernih vizualnih značilk ust ter združevanje akustične in vidne informacije v postopku razpoznavanja govora. Prav ti dve podpodročji zajemata večino problematike, s katero se srečujemo pri izgradnji bimodalnih sistemov za samodejno razpoznavanje govora.

1. Uvod

Najpomembnejši nosilec pri prenosu govorjenih sporočil predstavlja akustični signal. Za prenos govorne informacije je človek razvil tudi številne tehnološke pripomočke, ki omogočajo prenos akustičnega govornega signala na daljavo. Dokler je razmerje med signalom in šumom pri prenosu preko medija, tako pri neposredni medsebojni komunikaciji med ljudmi kot tudi na daljavo, dovolj veliko, je razpoznavanje govora, ki temelji le na sprejetem akustičnem signalu, uspešno. Pri postopku razpoznavanja govora pri ljudeh je poleg analize sprejetega dovolj kakovostnega akustičnega (slušnega) signala zelo pomembno tudi višjenivojsko znanje (jezik, tematika ipd.).

S psihološkimi raziskavami v zadnjih desetletjih so dokazali, da človek pri razpoznavanju govora uporablja tudi vizualno informacijo, kadar je ta na voljo (McGurk in MacDonald, 1976; Luettin, 1997). O pomembnosti vizualne informacije se lahko prepričamo tudi sami s preprosto primerjavo neposrednega osebnega pogovora s sogovornikom in pogovora po telefonu. Prav gotovo je pogovor, kjer sogovornika tudi vidimo, informacijsko precej bogatejši, hkrati pa ponavadi tudi prijetnejši zaradi socialnih razlogov, kar potrjuje tudi razvoj na področju video telekomunikacijskih tehnologij. Poleg zagotavljanja informacije o sogovornikovem razpoloženju in čustvenem stanju nam vidni stik pomaga tudi v procesu razpoznavanja govora, predvsem v primerih, kadar je akustični signal moten zaradi hrupnejšega okolja, npr. v avtomobilu, letalu, v tovarniškem okolju, pa tudi pri navzkrižnem govorjenju več govorcev, kar pogosto srečamo na javnih prostorih, v lokalih ipd. V takšnih primerih ljudje precej težje govorno komuniciramo, močno pa nam pomaga pozorno opazovanje gibanja govorčevega obraza, zlasti ust in spodnje čeljusti. Ljudje z okvarami sluha imajo zato pogosto izjemne sposobnosti branja ustnic oziroma vizualnega branja govora, še posebej pri izrazitejših gibih obraza in ustnic govorca ob poudarjeni artikulaciji.

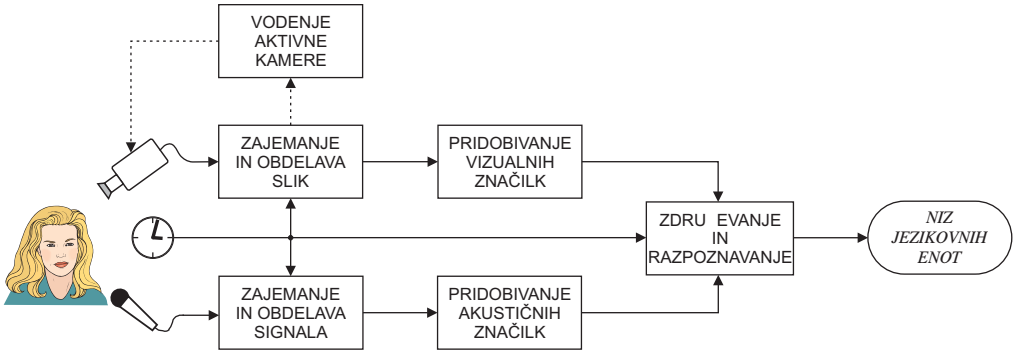
Zanimiva je trditev nekaterih ljudi, da dialoge v televizijskih oddajah bolje razumejo, če si nadenejo svoja očala. Televizijske oddaje s slabo sinhroniziranim zvočnim in slikovnim zapisom so za gledalce moteče. Podobno je s filmi in reklamnimi posnetki, sinhroniziranimi v tuj jezik, kjer nas zmede neujemanje slušne informacije z vidno. Kako močan je vpliv vizualne informacije na percepcijo govora, sta pokazala McGurk in MacDonald (1976) v članku s pomenljivim naslovom ,,Hearing lips and seeing voices“ , kjer opišeta presenetljiv učinek kombiniranja akustičnih (slušnih) in vizualnih vzbujanj pri ljudeh (t. i. ,,McGurkov efekt“). Kombinacijo video posnetka govorca, ki izgovarja zlog /ga/, s sinhroniziranim zvočnim posnetkom izgovorjenega zloga /ba/, človek razpozna kot zlog /da/, ki ni enak niti zvočnemu niti vidnemu vzbujanju. Če je preizkusna oseba izpostavljena le zvočnemu ali le vidnemu vzbujanju, je razpoznavanje posameznega zloga pravilno. Zanimivo je, da navzkrižna kombinacija obeh virov, kot smo jo opisali, ,,zmede“ preizkusno osebo tudi v primeru, če ta oseba pozna McGurkov efekt. McGurkov efekt je močnejši pri odraslih osebah kot pri otrocih, preizkusili in potrdili pa so ga tako na kratkih kombinacijah soglasnikov in samoglasnikov kot tudi na besedah in stavkih.

2. Bimodalno samodejno razpoznavanje govora

Samodejno razpoznavanje govora (angl. *Automatic Speech Recognition, ASR*) je precej razvito in uveljavljeno znanstveno-raziskovalno področje, katerega izsledke in tehnologije predvsem v zadnjem času že srečujemo v komercialno dostopnih izdelkih. Večinoma so ti sistemi močno občutljivi na kvaliteto vhodnega akustičnega signala in v hrupnem okolju ne delujejo dobro ali celo odpovedo, kar velja tudi za preprostejše sisteme, ki so sposobni razpoznavati le manjšo množico ločeno izgovorjenih ukazov.

Da bi izboljšali razpoznavanje govora pri motenem akustičnem signalu, si lahko tudi v sistemih za samodejno razpoznavanje govora pomagamo z vizualno (vidno) informacijo. Poleg akustične informacije, zajete z mikrofonom (slika 1a), v sistem sedaj vodimo še vizualno informacijo, predvsem o obrazu govorca, ki jo dobimo z video kamero (slika 1b). V tem primeru govorimo o *bimodalnem* (avdio-vizualnem) samodejnem razpoznavanju govora. Poleg podsistema za zajemanje in obdelavo vidnih

a)



b)

Slika 1: Samodejno razpoznavanje govora: unimodalni (a) in bimodalni (b) pristop.

podatkov je pri bimodalnem pristopu dodana še časovna sinhronizacija, ki je zelo pomembna predvsem v postopkih zajemanja vidnih in akustičnih podatkov ter v postopku razpoznavanja. Na sliki 1b je prikazana tudi možnost vodenja aktivne kamere (npr. barvna kamera EVI-D31 proizvajalca Sony), kar omogoča dinamično aktivno prilagajanje kota zasuka in nagiba ter zooma kamere med delovanjem sistema. Z aktivno kamero lahko torej „s pogledom sledimo“ obrazu govorca, kadar se ta premika. Na ta način skušamo doseči dobro izkoriščenost ločljivosti digitalnega slikovnega podsistema, saj lahko vplivamo na položaj in velikost slike obraza v zajetem zaporedju slik govorca.

Omenimo še, da je postopek razpoznavanja na sliki 1 simbolično zelo poenostavljeno prikazan le z enim samim blokom. Izvedbe razpoznavalnikov na področju samodejnega razpoznavanja govora so precej zapletene in lahko poleg akustičnega modela vključujejo tudi jezikovni model, ki predstavlja višjenivojsko znanje, kar na sliki ni prikazano. Prav tako zaradi enostavnosti na sliki 1 nismo prikazali postopka učenja razpoznavalnika.

Dodatek vizualne informacije, ki jo vodimo v sistem, v splošnem ne spreminja postopkov in načel tradicionalnih pristopov k razpoznavanju govora z uporabo akustične informacije. Za razpoznavanje se uporabljajo podobni postopki, ki največkrat temeljijo na prikritih Markovovih modelih ali nevronskih omrežjih (Rabiner in Juang, 1993; Jelinek, 1997). Pri bimodalnem pristopu se pojavi poseben problem združevanja (integracije) vizualne in akustične informacije, o katerem bo še govora v nadaljevanju. Kritično vprašanje, ki odpira široko področje problemov, pa je vsekakor pridobivanje relevantnih značilk iz slike (zaporedja slik) kamere. V nekaterih pristopih, znanih iz literature, so pridobivanje značilk močno poenostavili z barvanjem nekaterih delov obraza, predvsem ustnic, in uporabo na obraz prilepljenih značk - markerjev (Öhman, 1998; Barker in Berthommier, 1999), največkrat modre barve. Obarvane dele obraza in točkaste ali strukturne barvne značke je moč enostavno, učinkovito in zanesljivo izločati s preprostimi postopki obdelave slik. Takšni pristopi so sicer primerni za preizkusne laboratorijske razmere, niso pa ustrezni za izgradnjo splošnonamenskih avdio-vizualnih govornih vmesnikov med človekom in strojem. V nadaljevanju bomo zato opisali problematiko pridobivanja primernih vizualnih značilk brez uporabe značk ali kozmetičnih sredstev na obrazu govorca.

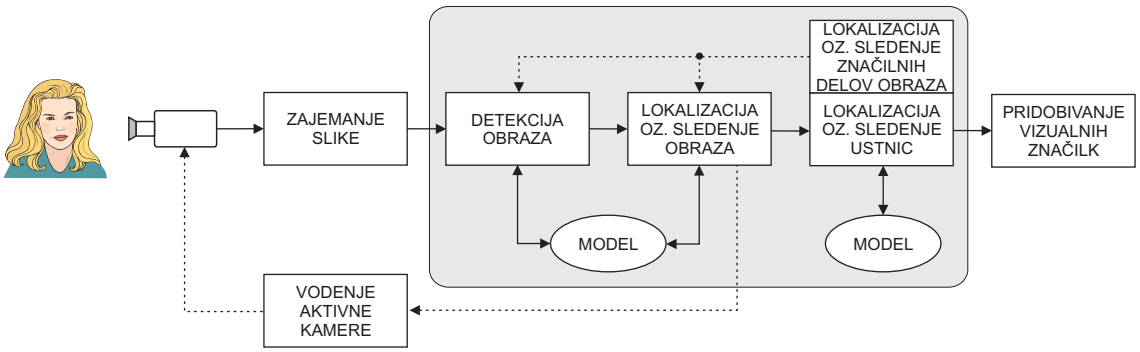
Glede na vrsto informacije, vključene v razpoznavanje govora pri ljudeh, ločimo (Matthews, 1998):

- Branje ustnic (angl. *lipreading*) je vizualna percepcija govora, ki temelji izključno na opazovanju artikulatornih gibov (ustnic) govorca brez poslušanja.
- Avdio-vizualna percepcija govora je percepcija govora, ki kombinira branje ustnic s poslušanjem.
- Branje govora (angl. *speechreading*) zajema širši obseg aktivnosti - gre za razumevanje govora, ki vključuje opazovanje govorčeve artikulacije, obraznih gibov in ročnih gest, lahko pa zajema tudi poslušanje.

Zgoraj navedene definicije s področja človeške percepcije se na področju bimodalnega samodejnega razpoznavanja govora uporabljajo precej ohlapno. Tako npr. naletimo na enačenje izrazov, kot so *avdio-vizualno razpoznavanje govora*, *samodejno branje ustnic* in *branje govora* (Hennecke et al., 1995).

2.1. Računalniška obdelava slik obrazov

Pri računalniški obdelavi slik obrazov na področju razpoznavanja govora nas zanimajo postopki s področja umetnega (računalniškega) vida, ki iz slike kamere pridobijo značilke, primerne za razpoznavanje. Računalniška



Slika 2: Računalniška obdelava slik obraza pri pridobivanju značilk.

obdelava slik človeških obrazov se je razvijala predvsem na področju samodejnega računalniškega razpoznavanja obrazov z namenom identifikacije oseb. Mnogo postopkov s tega področja je moč uporabiti ali prilagoditi tudi za pridobivanje značilk za vizualno podporo pri samodejnem razpoznavanju govora. Posebnost pri analizi obrazov za pridobivanje teh značilk je, da nas večinoma zanima le spodnji del obraza, to je področje ust z okolico in spodnja čeljust. V splošnem je celoten postopek računalniške obdelave slik obraza sestavljen iz naslednjih postopkovnih skupin (slika 2):

- detekcija (ugotavljanje prisotnosti) obraza v sliki,
- določanje lokacije obraza na sliki (inicializacija sledenja) oziroma sledenje področja obraza skozi zaporedje slik,
- lokalizacija in sledenje značilnih delov obraza, predvsem ustnic, znotraj obraza,
- pridobivanje vizualnih značilk, primernih za razpoznavanje.

Meja med posameznimi postopkovnimi skupinami ni vedno izrazita, kar je na sliki 2 ponazorjeno z osenčenim področjem, ki zajema posamezne postopkovne bloke. Nemalokrat so npr. med seboj močno povezani ali celo združeni moduli za detekcijo in lokalizacijo oz. sledenje področja obraza in njegovih značilnih delov. Tako lahko modula za detekcijo in lokalizacijo obraza generirata hipoteze o prisotnosti in lokaciji obraza na sliki, ki jih sprejmemo ali zavržemo na podlagi analize obraznega področja z iskanjem značilnih delov, kot so obrvi, oči, nosnice in usta. Na sliki 2 je to nakazano s črtkano povezavo med blokom, ki predstavlja lokalizacijo oz. sledenje značilnih delov obraza, ter blokoma detekcije in lokalizacije oz. sledenja obraza. Vrstni red modulov za lokalizacijo oz. sledenje obraza in značilnih delov obraza je lahko takšen, kot smo ga navedli, ali pa obraten, saj lahko natančno lego obraza določimo prav na podlagi detekcije in lokalizacije določenih značilnih obraznih področij, predvsem oči.

Za določanje področij v sliki, ki so kandidati za obraz, pogosto uporabljamo barvno informacijo. Odtenek (angl. *hue*) barve kože je namreč za različne osebe, ki lahko celo pripadajo različnim človeškim rasam, zelo podoben. S pomočjo barvnega modela kože lahko pri nadzorovani barvi osvetlitve hitro in dokaj natančno opravimo osnovno segmentacijo slike in tako na sliki poiščemo področja kožne barve. Metoda je neodvisna od orientacije obraza na sliki. Kot model lahko uporabimo dvodimenzionalni Gaussov model (Stiefelhagen et al., 1997; Yang in Ahuja, 1998) v enem od barvnih prostorov ali model utežene vsote več (dveh) Gaussovih porazdelitev (angl. *Gaussian mixture model, GMM*), npr. Chen (2001). Za zmanjšanje odvisnosti od osvetljenosti obraza uporabljamo različne postopke za odstranjevanje informacije o intenziteti, npr. v RGB barvnem prostoru z normiranjem RGB vrednosti posameznega slikovnega elementa, tako da je vsota normiranih barvnih komponent enaka 1. V ravnini normiranih (r,g) komponent nato zgradimo barvni model kože (Stiefelhagen et al., 1997). Podoben pristop je pretvorba v HSV ali HLS barvni prostor, ki sta bližje človeški percepciji barv, ter izgradnja primernega barvnega modela v (H,S) prostoru. Rezultat segmentacije slike z barvnim modelom kože so področja slike, ki z veliko verjetnostjo pripadajo barvi kože. Ta področja seveda ne pripadajo nujno sliki obraza, temveč tudi rokam, vratu, pa tudi predmetom iz ozadja, ki so koži podobne barve (npr. pohištvo v naravni barvi nekaterih vrst lesa). Dobljena segmentirana področja slike (hipoteze) predstavimo v obliki binarne maske. Ta področja moramo sedaj analizirati ter jih sprejeti kot tista, ki pripadajo sliki obraza, ali pa jih zavreči. Pred tem jih običajno še obdelamo s primernim zaporedjem postopkov, kot so npr. morfološke operacije razjedanje ali erozija (angl. *erosion*) in razširjanje ali dilatacija (*dilation*) oziroma njuni kombinaciji odpiranje (*opening*) in zapiranje (*closing*), analiza povezanih področij in zapolnjevanje s preplavljanjem ter brisanje področij s površino, manjšo od izbranega praga. Ko je področje, ki pripada obrazu, določeno, običajno opišemo lokacijo obraza z elipso, ki se prilega temu področju. Parametre elipse določimo na podlagi izračuna momentov področja (Sobottka in Pitas, 1996).

Detekcija in lokalizacija obraza na statičnih slikah predstavljata v splošnem bolj zapleten problem kot sledenje obraza oziroma značilnih obraznih delov v zaporedju slik (video sekvenci). Pri zaporedju slik lahko uporabimo informacijo o gibanju obraza v prizoru ali gibanju značilnih delov obraza znotraj obraznega področja (npr. pri govoru), pomagamo pa si lahko tudi z odštevanjem slike ozadja, ki jo skozi zaporedje slik na primeren način stalno obnavljamo in s tem odstranimo (počasnejše) spremembe v ozadju. Pri

sledenju je z informacijo iz prejšnjih slik zaporedja mogoče omejiti področje iskanja obraza oziroma njegovih značilnih delov znotraj posamezne slike v zaporedju. Informacijo o lokaciji obraza na trenutni sliki in njegovem gibanju skozi zaporedje slik lahko uporabimo tudi za vodenje aktivne kamere (slika 2) ter na ta način vzdržujemo primeren položaj in velikost (slike) obraza skozi zaporedje slik kljub premikanju govorca.

2.1.1. Lokalizacija značilnih delov obraza ter pridobivanje značilk

Pri lokalizaciji in sledenju značilnih delov obraza največkrat uporabimo zaporedje (sekvenco) slik obraza, posneto od spredaj. Ker tudi človek med pogovorom ponavadi gleda sogovornikov obraz s sprednje strani (frontalno), večina znanih izvedb sistemov za vizualno branje ustnic temelji na takšnem pogledu. Ustreznost takšne izbire pogleda potrjuje tudi dejstvo, da je človek sposoben izkoriščati vizualno govorno informacijo iz frontalnih dvodimenzionalnih video posnetkov, npr. pri spremljanju televizijskega dnevnika.

Preslikava zaporedja slik v zaporedje vrednosti vizualnih značilk, ki bi izboljšale rezultat razpoznavanja govora, predstavlja eno temeljnih raziskovalnih vprašanj pri razvoju sistemov za vizualno podporo akustičnim razpoznavalnikom govora. Čeprav nas za pridobivanje primernih vizualnih značilk zanima prav področje ust, pa v sistemih za vizualno branje ustnic včasih lokaliziramo in sledimo več značilnih področij obraza, npr. oči, obrvi, nosnice, in ne le ust. Pri nekaterih postopkih je namreč laže lokalizirati in slediti druga značilna področja obraza, nato pa z upoštevanjem geometrijskih razmerij določiti še položaj področja ust. Poleg tega je z upoštevanjem medsebojnih položajev značilnih področij mogoče povečati robustnost postopkov sledenja (Stiefelhagen et al., 1997). Pri frontalnih posnetkih si lahko pomagamo tudi z upoštevanjem zrcalne simetrije obraza okrog navpične osi. Postopki za lokalizacijo ter sledenje značilnih delov obraza so ponavadi primerno prilagojeni znani postopki s področja računalniškega vida, npr. koreliranje z modelom, uporaba gradientnih (sivih) slik in projekcij vzdolž osi, Houghova transformacija, valčna transformacija, segmentacija z aktivnimi krivuljami in deformabilnimi predlogami, uporaba kompleksnih prilagodljivih (aktivnih) modelov značilnih obraznih področij, predvsem ust in oči itd.

Ko je področje ust na sliki lokalizirano, sledi postopek izločanja vrednosti značilk ust za razpoznavanje. Pojem značilka (angl. *feature*) lahko označuje skoraj poljuben vidik predstavitve. Tako lahko za opis ustnega področja uporabimo geometrijske značilke, kakršne so npr. položaj ustnih kotičkov, zunanji in notranji obris zgornje in spodnje ustnice, širina ust, odprtost ust, površina in oblika odprtja ust ipd. Nekateri postopki uporabljajo kot značilke kar vrednosti slikovnih elementov sive slike ust, ponavadi pri zmanjšani ločljivosti, kjer vključijo še analizo glavnih komponent (angl. *principal component analysis, PCA*) za dodatno zmanjšanje dimenzije (preslikanega) vektorja značilk. Zelo uspešni so tudi kombinirani postopki, ki zajemajo tako geometrijske značilke za opis oblike kot tudi slikovne značilke v obliki vrednosti slikovnih elementov. Takšne kombinirane pristope je moč uporabiti tudi za opis področja znotraj ust. V splošnem namreč velja, da je največ vizualne informacije vsebovane v obrisu (konturi) ustnic, pomembno pa je tudi področje znotraj ust, kjer nas zanima zlasti vidnost zob in jezika (Luettin, 1997; Summerfield et al., 1989). Za opis oblike so v literaturi znani tudi modeli, ki jih prilagajamo trenutni obliki in/ali sivinskim nivojem slike ustnic, npr. Matthews et al. (2002). Prilagajanje modela pomeni spreminjanje njegovih parametrov, ki v tem primeru lahko služijo kot značilke za razpoznavanje, ali pa značilke izpeljemo iz modela (npr. geometrijske značilke pri modelih oblike ustnic).

Pri omenjenih postopkih lahko uporabimo statične ali dinamične značilke ali kombinacijo obeh. Prav dinamične značilke za opis časovnih sprememb položaja ustnic po nekaterih ugotovitvah (Bregler in Konig, 1994) nosijo pomembnejšo informacijo od opisa absolutnega položaja ustnic.

2.1.2. Problematika računalniške obdelave slik obrazov

Za človeka predstavljata detekcija in razpoznavanje obrazov zelo preprosto vsakodnevno nalogo. Iz obrazov drugih ljudi smo zmožni razbrati njihove občutke, razpoloženje, čustva, pa tudi raznovrstne, včasih zelo majhne spremembe. Obstajata dve hipotezi o človeškem zaznavanju in razpoznavanju obrazov: človeški vizualni sistem razpoznava obraze kot celoto ali (druga hipoteza) kot zbirko značilnih delov ali značilk, ki sestavljajo obraz (Bala et al., 1997). Podobnost s tema dvema hipotezama najdemo tudi pri samodejnih sistemih za lokalizacijo oziroma sledenje obrazov, kar smo že omenjali pri vrstnem redu modulov za detekcijo obraza in njegovih značilnih delov. Vendar je v nasprotju s človeškim zaznavanjem splošnejša rešitev nalog detekcije in lokalizacije obraza precej zapleten problem računalniškega vida. Pri postopkih računalniške obdelave slik človeških obrazov, predvsem pri lokalizaciji značilnih delov obraza, naletimo na vrsto težav, ki so neposredno povezane z množico pojavnih oblik, v kakršnih lahko ti obrazi nastopajo. Obrazi so lahko močno različni, a imajo vsi enako osnovno strukturo. Kljub temu splošne formalne definicije obraza, uporabne za implementacijo v računalniških postopkih, ne znamo zapisati. Postopki računalniške obdelave slik obrazov zato delujejo s pomočjo različnih predpostavk. Včasih so lahko že različni izrazi na obrazu vzrok za velike razlike v videzu istega obraza. Te razlike se odražajo v slikah in lahko povzročijo poslabšanje delovanja ali celo odpoved postopkov računalniške obdelave, saj nekatere predpostavke ne veljajo več. Težave pri delovanju postopkov računalniške obdelave (barvnih) slik obrazov na področju bimodalnih sistemov za razpoznavanje govora izhajajo predvsem iz naslednjih razlogov:

- osvetlitev - neenakomerna osvetlitev obraza, sence in odsevi (npr. od stekel očal), razlike med umetno in naravno (dnevno) svetlobo, barva umetne svetlobe
- obrazna poraščenost pri moških - brada in brki ovirajo predvsem postopke lokalizacije in sledenja ustnic
- dekorativna kozmetika pri ženskah - predvsem uporaba nenaravnih in močnejših barv kozmetičnih

sredstev (make-up)

- lasje - neobičajne barve las in močno nesimetrične pričeske, delna zakritost obraza z lasmi
- korekcijska in predvsem sončna očala, obrazni nakit (npr. uhani, prilepljeni okraski), šali, pokrivala ipd.
- delna zakritost obraza zaradi drugih razlogov - npr. zakrivanje ust ali podpiranje glave z roko ipd.

2.2. Združevanje akustične in vizualne informacije

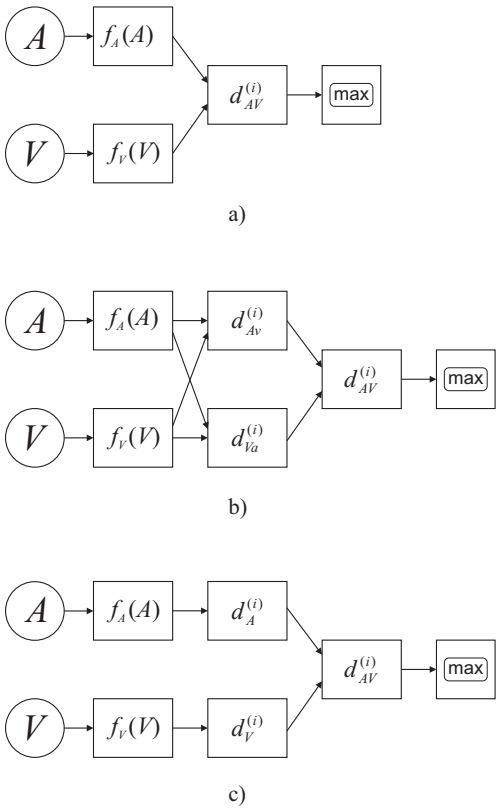
Kot smo prikazali na sliki 1b, potekata obdelavi akustičnih in vizualnih podatkov v bimodalnem sistemu ločeno. V procesu razpoznavanja je potrebno združiti pridobljene obdelane akustične in vizualne podatke. Združevanje akustične in vizualne informacije je močno povezano z zgradbo razpoznavalnika. Pri združevanju nas predvsem zanimata način in nivo, na katerem združimo informaciji obeh virov. Glede na nivo združevanja ločimo dve skrajni možnosti: *zgodnji* (angl. *early*), slika 3a, in *pozni* (angl. *late*) model, slika 3c. Na sliki 3 so z A in V označeni akustični oziroma vizualni podatki. Obdelava teh podatkov je ponazorjena z bloki $f_A(A)$ in $f_V(V)$, funkcije $d^{(i)}_{AV}$, $d^{(i)}_{Av}$, $d^{(i)}_{Va}$, $d^{(i)}_{A}$ in $d^{(i)}_{V}$ pa so diskriminantne oziroma odločitvene funkcije posameznih razredov i. Izhodni razred je moč določiti z iskanjem ekstrema (maksimuma) teh funkcij.

Pri zgodnjem modelu združevanja informacije sestavimo (zlepimo) vektorja akustičnih in vizualnih značilk v en sam daljši vhodni vektor, ki ga vodimo v razpoznavalnik. Pri poznem modelu imamo dva povsem ločena razpoznavalnika, enega za akustični in drugega za vizualni vektor značilk, katerih izhoda (izračunane verjetnosti razredov) moramo na primeren način kombinirati. Če informacijo iz obeh virov združujemo na kakem drugem, vmesnem nivoju, govorimo o *vmesnem* (angl. *intermediate*) modelu združevanja, kar je simbolično prikazano na sliki 3b z navzkrižnima povezavama med obema vejama. Znanih je le nekaj arhitektur vmesnega združevanja (Hennecke et al., 1995).

Težave, na katere naletimo pri postopkih združevanja akustične in vizualne informacije, so predvsem (Verma et al., 1999):

- različno število razredov, med katerimi razlikujemo pri akustičnih in vizualnih podatkih (različno število fonemov in vizemov),
- različna vzorčevalna frekvenca video in avdio signala (potreba po interpolaciji video podatkov),
- časovni zamik med avdio in video signalom, ki izhaja iz kompleksnosti artikulatornih pojavov,
- razlike v dinamičnih območjih avdio in video značilk.

Zgodnji model predstavlja (teoretično) najbolj splošno metodo združevanja informacije. Velja namreč, da bi pri pravilni izbiri arhitekture sistema za razpoznavanje in metode učenja razpoznavalnik, zgrajen po zgodnjem modelu, moral delovati vsaj tako dobro kot sistem, ki združuje informacijo na kasnejši stopnji (Hennecke et al., 1995). Če



Slika 3: Zgodnji (a), vmesni (b) in pozni (c) model združevanja akustičnih in vizualnih podatkov.

bi se npr. pozni model izkazal za optimalnega pri avdiovizualnem branju govora, bi se takšen razpoznavalnik torej naučil izvajati pozno združevanje. Razpoznavalnik mora biti sposoben učenja avtomatskega uteževanja informacije glede na njeno točnost in jezikovno umestnost. Ker sta sinhrona vektorja akustičnih in vizualnih značilk pri zgodnjem modelu med seboj združena že na vhodu razpoznavalnika, mora biti takšen razpoznavalnik sposoben tudi učenja relativnih časovnih razlik med obema viroma. Pri tem je ključnega pomena dobra sinhronizacija virov, ki ni vedno zagotovljena. Razpoznavalniki, ki uporabljajo zgodnji model združevanja, imajo več parametrov kot tisti, ki združujejo pozno, zato potrebujejo tudi več učnih podatkov. Da bi npr. dosegli pravilno uteževanje med obema viroma informacije v odvisnosti od nivoja šuma v akustičnem signalu, moramo razpoznavalnik učiti preko celotnega območja šumnih nivojev, ki lahko nastopijo v fazi razpoznavanja, kar še dodatno močno poveča učno množico podatkov.

Pri razpoznavanju s poznim modelom združevanja (slika 3c) je sistem za razpoznavanje zgrajen iz dveh ločenih razpoznavalnikov; v prvega vodimo vektor akustičnih, v drugega pa vektor vizualnih značilk. S tem lahko rešimo nekatere prej naštete težave združevanja podatkov. Vrednosti diskriminantnih (npr. verjetnostnih) funkcij se določajo ločeno za akustične in vizualne značilke v vsakem od obeh podsistemov, nato pa na primeren način kombinirajo v en rezultat. Najpreprostejši način takšnega kombiniranja je s predpostavko statistične

neodvisnosti obeh kanalov, kjer preprosto zmnožimo obe izhodni verjetnosti razredov. Zaradi ločenega delovanja razpoznavalnikov sinhronizacija med akustičnim in vizualnim kanalom ni kritična, arhitektura obeh razpoznavalnikov pa je lahko različna. Model poznega združevanja zato omogoča nadgradnjo že obstoječih sistemov samodejnega akustičnega razpoznavanja govora. Pozni model združevanja je primeren za uporabo le v razpoznavalnikih, kjer nastopa manjša množica izgovorjenih enot (angl. *utterances*). Za vsako od možnih enot je namreč potrebno določiti obe pripadajoči verjetnosti (izhod akustičnega in vizualnega podsistema). Model poznega združevanja je zato večinoma uporabljen na fonemskem nivoju, na katerem so nadgrajeni še besedni in stavčni razpoznavalniki.

Vmesni model združevanja akustične in vizualne informacije je močno povezan z arhitekturo celotnega razpoznavalnika, ki mora omogočati vpliv enega (pod)razpoznavalnika na drugega, kar lahko ohrani nekatere prednosti tako zgodnjega kot poznega modela oziroma odstrani njune slabosti. V literaturi napogosteje naletimo na realizacije zgodnjega ali poznega združevanja, pristopi z vmesnim modelom pa so precej redki.

2.2.1. Akustično-vizualni razpoznavalniki

Osnovna naloga akustično-vizualnega (tudi avdio-vizualnega) razpoznavalnika govora je predvsem primerno združevanje akustične in vizualne (lahko tudi konfliktne) informacije na tak način, da kar najbolj zmanjšamo napake pri razpoznavanju. Pri akustično-vizualnem razpoznavanju uporabljamo tehnike, znane iz akustičnega razpoznavanja govora (Rabiner in Juang, 1993; Jelinek, 1997), najpogosteje so to prikriti Markovovi modeli (HMM) in nevronska omrežja (NN).

Uporaba zgodnjega modela pri HMM razpoznavalniku lahko dá slabše rezultate kot pozni model združevanja (Hennecke et al., 1995), kar je v nasprotju z opisano tezo o splošni prednosti zgodnjega združevanja. Razlog je najverjetneje v različni (vzorčevalni) hitrosti (angl. *rate*) akustične in vizualne informacije, česar se prikriti Markovovi modeli niso sposobni (pravilno) naučiti. Pri poznem združevanju informacije s prikritimi Markovovimi modeli gre za razpoznavanje z dvema vzporednima HMM-jema, ki se lahko vsak posebej prilagodita na hitrost svojega (akustičnega oziroma vizualnega) kanala. Izhoda obeh razpoznavalnikov s primerno metodo kombiniramo v končni rezultat.

Kot primer vmesnega modela združevanja omenimo *Boltzmannovo zadrgo*, ki je sestavljena iz dveh linearnih verig nevronskih enot z vmesnimi križnimi povezavami. Vsaka od obeh verig lahko vsebuje različno število enot, več za akustično informacijo in manj za vizualno. Boltzmannovo zadrgo nato pretvorimo v polno povezan HMM, ki ga uporabimo za razpoznavanje. Prvi rezultati takšnega vmesnega združevanja kažejo izboljšanje tako v primerjavi z zgodnjim kot poznim modelom združevanja (Hennecke et al., 1995).

Za razliko od razpoznavanja s HMM je pri razpoznavalnikih z nevronskimi omrežji težko doseči neodvisnost od hitrosti govora, učenje pa je počasno. Vendar pa so nevronska omrežja sposobna dobro posploševati z manj predpostavkami o podatkih, če je le na voljo dovolj velika učna množica. Tudi združevanje je ponavadi dokaj enostavno ne glede na izbrani model združevanja. Pri poznem modelu imamo dve ločeni nevronski omrežji s podobno arhitekturo, ki imata v splošnem različno število skritih enot. V prvo omrežje vodimo vektor akustičnih, v drugo pa vektor vizualnih značilk. Navzkrižnih povezav med omrežjema pri poznem modelu ni, končni rezultat pa dobimo s kombiniranjem izhodnih vrednosti unimodalnih diskriminantnih funkcij obeh omrežij. Če namesto kombiniranja izhodnih vrednosti napeljemo navzkrižne povezave, dobimo vmesni model združevanja. Nivo (stopnjo) združevanja določa število navzkrižnih povezav in plasti, kjer se te povezave pojavljajo. Izhodna plast je v teh primerih le ena, diskriminantna funkcija na izhodu pa je že bimodalna. V skrajnem primeru (zgodnji model) dobimo eno samo povsem povezano nevronsko omrežje.

2.3. Avdio-vizualne podatkovne zbirke

Pri izgradnji sistemov za samodejno avdio-vizualno razpoznavanje govora potrebujemo primerno zbirko avdio-vizualnih posnetkov tako za učenje kot tudi preizkušanje zgrajenih sistemov. Pri vrednotenju delovanja avdio-vizualnih razpoznavalnikov in njihovih podsistemov bi potrebovali tudi standardne zbirke, s katerimi bi lahko primerjali uspešnost pristopov različnih raziskovalnih skupin. Ena od težav pri mednarodnem primerjanju rezultatov in sprejetju standardnih preizkusnih avdio-vizualnih zbirk je prav gotovo uporabljeni jezik, saj postopki samodejnega razpoznavanja govora praviloma vključujejo tudi posebnosti jezika. Posamezne raziskovalne skupine večinoma posnamejo svoje avdio-vizualne podatkovne zbirke.

Problematika avdio-vizualnih podatkovnih zbirk je predvsem tehnične narave in je povezana predvsem s samim pridobivanjem (snemanjem) podatkov, njihovim shranjevanjem na primerne medije ter razmnoževanjem in distribucijo kopij zbirk. Zbirke so večinoma v digitalni obliki, zato so primerne za shranjevanje na računalniških diskih ali arhivirnih medijih. Digitalna oblika zagotavlja tudi kvalitetno reprodukcijo in možnost razmnoževanja podatkov brez izgube kvalitete. Zaradi velike količine podatkov, k čemur seveda največ prispevajo barvni video posnetki, so avdio-vizualne zbirke shranjene praviloma v komprimirani obliki (npr. v DV ali MPEG-2 formatu). Posnetki so ponavadi shranjeni in distribuirani na več medijih večje kapacitete (npr. DVD-RAM diski s kapaciteto 5,2GB ali 9,4GB).

Pri akustičnih podatkih zbirke nas zanima predvsem vzorčevalna frekvenca, število kvantizacijskih nivojev posameznega vzorca (npr. 16-bitni vzorci), način kvantizacije ter razmerje med signalom in šumom. Kadar so avdio in video posnetki v ločenih datotekah, beležimo sinhronizacijske podatke za oba dela zbirke v posebne datoteke (Potamianos et al., 1997). Koristen je tudi podatek o maksimalnem neujemanju (desinhronizaciji) med avdio in video posnetkom. Video sekvenca je lahko posneta črnobelo (sive slike) ali barvno, pri čemer so možne različne strukture (formati) vzorčenja luminance in obeh barvnih

(krominančnih) razlik (npr. 4:1:1 za NTSC DV in 4:2:0 za PAL DV). Število slik v sekundi je lahko različno, v Evropi ponavadi 25 (PAL televizijski sistem), v Združenih državah Amerike pa 30 (NTSC televizijski sistem) slik v sekundi. Vsaka od slik v video sekvenci ima določeno število slikovnih elementov (npr. 720 krat 576 slikovnih elementov), kar določa ločljivost podrobnosti na posamezni sliki. Pri video sekvenci podatkovnih zbirk nas poleg naštetih tehničnih podatkov zanimajo še pogoji osvetlitve, pri katerih je bila zbirka posneta, vsebina slik video sekvenc (zgornji del telesa in glava, obraz, samo spodnji del obraza), morebitna uporaba barvnih značk ali barvil na obrazu ipd.

Splošni podatki za avdio-vizualne zbirke so še: število in sestava govorcev (predvsem spol), vsebina in sestava govornega materiala, časovni intervali, v katerih so bili posnetki pridobljeni, morebitno označevanje zbirke, ki je ponavadi izvedeno ročno ali polavtomatsko ipd.

Avdio-vizualne podatkovne zbirke, ki jih je mogoče kupiti za raziskovalne namene, so danes še precej redke. Ena od njih je zbirka XM2VTSDB - „Extended M2VTS Database" (Messer et al., 1999), katere del imamo tudi v Laboratoriju za umetno zaznavanje, sisteme in kibernetiko na Fakulteti za elektrotehniko v Ljubljani. Gre za razmeroma obsežno multimodalno podatkovno zbirko, ki je v prvi vrsti namenjena učenju in preizkušanju multimodalnih sistemov za identifikacijo (verifikacijo) oseb, pa tudi za raziskave na področju bimodalnih sistemov za samodejno razpoznavanje govora, predvsem za preizkuse postopkov sledenja ustnic. Zbirko sestavljajo digitalni posnetki 295 različnih oseb obeh spolov. Posnetki so bili pridobljeni v štirih ločenih sejah, enakomerno razporejenih v obdobju petih mesecev. Poleg barvnih video posnetkov (z zvokom) sestavljajo zbirko še video posnetki počasnega obračanja glave (levo-desno in navzgor-navzdol), statični frontalni posnetki (slike) z enakomerno in neenakomerno osvetlitvijo obraza, statični stranski (profilni) posnetki in tridimenzionalni VRML model glave vsake od sodelujočih oseb. Podrobneje si bomo ogledali le avdio-vizualni del.

Barvni video posnetki so shranjeni na DVD-RAM medijih v DV kodiranem AVI formatu s 25 slikami v sekundi, strukturo vzorčenja 4:2:0 in stalnim razmerjem stiskanja podatkov 5:1. Vsaka slika (angl. *frame*) v sekvenci ima 720 krat 576 slikovnih elementov. Vzorčevalna frekvenca zvočnega signala je 32kHz, zvočni vzorci pa so 16-bitni. Preizkusne osebe izgovarjajo tri različne nize v angleškem jeziku, pri čemer je tretji niz sestavljen tako, da je glasovno uravnotežen:

1. „0 1 2 3 4 5 6 7 8 9"
2. „5 0 6 9 2 8 1 3 7 4"
3. „Joe took fathers green shoe bench out"

Vsaka od 295 oseb izgovori vsakega od treh nizov po enkrat v vsaki od štirih sej. Avdio-vizualni del zbirke torej vsebuje 3540 posnetih sekvenc, po 12 na vsako od sodelujočih oseb.

2.4. Področja uporabe

Pri razvoju inteligentnih vmesnikov med človekom in računalnikom ima pomembno vlogo avdio-vizualna percepcija govora, pa tudi razpoznavanje človeških kretenj in gest. Sodobni osebni računalniki so že večinoma opremljeni z multimedijsko avdio in video opremo, cenovno dostopne pa so tudi kamere, ki so praviloma nameščene na vrhu monitorja in „gledajo" uporabnika. Tovrstno opremo je mogoče s primerno programsko podporo uporabiti tudi kot bimodalni govorni vmesnik med človekom in računalnikom. Uporaba vizualne informacije v sistemih za samodejno razpoznavanje govora lahko močno pripomore k robustnosti delovanja govornih vmesnikov med človekom in strojem, predvsem v hrupnejših okoljih. Naštejmo le nekaj pomembnejših primerov uporabe:

Pisarne in uradi: Značilno za pisarne in urade je razmeroma hrupno okolje, npr. zaradi hkratnega govora več govorcev (pogovori s strankami, telefonski pogovori), pa tudi hrupa naprav (tiskalniki, fotokopirni stroji ipd.). Bimodalne avdio-vizualne govorne vmesnike bi bilo moč uporabiti predvsem pri povečanju robustnosti samodejnih sistemov za narekovanje teksta strojem (računalnikom), pa tudi pri robustnih govornih vmesnikih za ukazovanje inteligentnim pisarniškim napravam (fotokopirni stroji, komunikacijske in faksimilne naprave ipd.).

Pomoč osebam z okvaro sluha in invalidnim osebam: Kot pomoč osebam z okvaro sluha omenimo predvsem samodejne prepise in podnaslavljanje televizijskih oddaj, kjer je pogosto na voljo tudi slika govorca (npr. TV dnevnik). Ker si osebe z okvaro sluha učinkovito pomagajo z branjem ustnic govorca, jim lahko pomagamo tudi z računalniško animiranimi sintetičnimi obrazi. Za realistično animacijo so nujno potrebni rezultati raziskav s področja vizualnega branja govora. Uporaba inteligentnih vmesnikov, zasnovanih na postopkih avdio-vizualne percepcije govora in razpoznavanju človeških kretenj, v različnih inteligentnih hišnih napravah in aparatih lahko močno olajša življenje invalidnim osebam.

Varnostni sistemi in nadzor dostopa: Identifikacija oseb izključno na osnovi akustične govorjene informacije kot tudi izključno na osnovi vizualne informacije (razpoznavanje obraza) ne zagotavljata dovolj zanesljive metode za ugotavljanje istovetnosti osebe. Postopki, ki jih uporabljamo pri avdio-vizualnem branju govora, bi lahko omogočili izvedbo zanesljivejšega sistema za identifikacijo oseb na osnovi bimodalne avdio-vizualne identifikacije.

Uporaba v vozilih: Vgrajeni vmesniki za govorno upravljanje z radijskim sprejemnikom ali mobilnim telefonskim aparatom v avtomobilu danes niso več posebnost. Z majhno video kamero, nameščeno pri zgornjem robu vetrobranskega stekla nad voznikom ali na armaturni plošči, bi bilo mogoče robustnost in uporabnost teh vmesnikov v hrupni notranjosti avtomobila še izboljšati. Kamero je moč uporabiti tudi v druge namene, tako npr. že deluje prototip sistema za detekcijo prekomerne utrujenosti voznika.

Telekonference: Pri prenosu telekonferenc je pomemben tako prenos slikovne kot zvočne informacije. Iz frontalne slike govorca je moč izločati obrazne značilke in jih skupaj z značilkami, izločenimi iz akustične informacije, uporabiti za samodejni prepis (transkripcijo) govorjenega sporočila. Postopke iskanja, lokalizacije in sledenja obraza z video kamero, ki jih uporabljamo pri izvedbi samodejnih sistemov branja ustnic ali govora, je moč uporabiti tudi za av-

tomatizacijo telekonferenc.

Videotelefonija: Tudi pri videotelefoniji je mogoče dograditi podsistem za samodejni prepis govorjenih prenašanih sporočil. Prepisi telefonskih pogovorov so pogosto potrebni predvsem v poslovnem svetu. V prihodnosti lahko pričakujemo tudi avtomatske prevajalnike avdio-vizualnih govorjenih telefonskih sporočil, kjer bodo izločene vizualne obrazne značilke nujno potrebne za sintezo video sekvence govorečih obrazov.

3. Zaključek

V prispevku smo predstavili kratek pregled področja bimodalnega avdio-vizualnega razpoznavanja govora ter opisali problematiko posameznih korakov izgradnje bimodalnih sistemov. Za zanesljivo in robustno delovanje zgrajenega sistema je ključnega pomena prav podsistem za obdelavo vizualne informacije in pridobivanje vizualnih značilk. Najtežavnejši del izgradnje takega podsistema je robustna izvedba obdelave zaporedja slik govorčevega obraza, še posebej sledenje ustnic za poljubnega govorca. Od rezultatov teh postopkov je neposredno odvisen tudi uspeh postopka pridobivanja značilk, ki jih uporabimo v procesu avdio-vizualnega samodejnega razpoznavanja govora.

Tudi v Laboratoriju za umetno zaznavanje, sisteme in kibernetiko na Fakulteti za elektrotehniko v Ljubljani se ukvarjamo z razvojem podsistema za zajemanje in obdelavo vizualnih podatkov ter pridobivanje vizualnih značilk govorcev. Takšen podsistem bo služil kot osnova za raziskovalno delo na področju bimodalnega samodejnega razpoznavanja slovenskega govora ter razvoja inteligentnih vmesnikov med človekom in strojem.

4. Literatura

L.-P. Bala, K. Talmi in J. Liu. 1997. Automatic Detection and Tracking of Faces and Facial Features in Video Sequences. V: *Proceedings of Picture Coding Symposium (PCS '97)*, str. 251–256, Berlin.

J. P. Barker in F. Berthommier. 1999. Evidence of Correlation Between Acoustic and Visual Features of Speech. V: *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS99)*, str. 199–202, San Francisco.

C. Bregler in Y. Konig. 1994. "Eigenlips" for Robust Speech Recognition. V: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, zvezek 2, str. 669–672, Adelaide.

T. Chen. 2001. Audiovisual Speech Processing - Lip Reading and Lip Synchronization. *IEEE Signal Processing Magazine*, 18(1):9–21.

M. E. Hennecke, D. G. Stork in K. V. Prasad. 1995. Visionary Speech: Looking Ahead to Practical Speechreading Systems. V: D. G. Stork in M. E. Hennecke, ur., *Speechreading by Humans and Machines*, zvezek 150, NATO ASI Series, Series F: Computer and System Sciences, str. 331–349. Springer Verlag, Berlin.

F. Jelinek. 1997. *Statistical Methods for Speech Recognition*. MIT Press.

J. Luettin. 1997. *Visual Speech and Speaker Recognition*. Doktorsko delo, University of Sheffield.

I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox in R. Harvey. 2002. Extraction of Visual Features for Lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:198–213.

I. Matthews. 1998. *Features for Audio-Visual Speech Recognition*. Doktorsko delo, School of Information Systems, University of East Anglia.

H. McGurk in J. MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264:746–748.

K. Messer, J. Matas, J. Kittler, J. Luettin in G. Maitre. 1999. XM2VTSDB: The Extended M2VTS Database. V: R. Chellapa, ur., *Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99)*, str. 72–77, Washington.

T. Öhman. 1998. An audio-visual speech database and automatic measurements of visual speech. Quarterly Progress and Status Report 1-2/1998, Royal Institute of Technology (KTH), Stockholm.

G. Potamianos, E. Cosatto, H. P. Graf in D. B. Roe. 1997. Speaker Independent Audio-Visual Database for Bimodal ASR. V: *Proceedings of the European Tutorial and Research Workshop on Audio-Visual Speech Processing (AVSP'97)*, str. 65–68, Rhodes.

L. Rabiner in B.-H. Juang. 1993. *Fundamentals of Speech Recognition*. Prentice Hall.

K. Sobottka in I. Pitas. 1996. Localization of Facial Regions and Features in Color Images. V: *4th Open Russian-German Workshop: Pattern Recognition and Image Analysis*, str. 134–138, Valday.

R. Stiefelhagen, U. Meier in J. Yang. 1997. Real-Time Lip-Tracking for Lipreading. V: *Proceedings of the 5th European Conference on Speech Communication and Technology EUROSPEECH'97*, zvezek 4, str. 2007–2010, Rhodes.

Q. Summerfield, A. MacLeod, M. McGrath in M. Brooke. 1989. Lips, teeth and the benefits of lipreading. V: A. W. Young in H. D. Ellis, ur., *Handbook of Research on Face Processing*, str. 223–233. Elsevier Science Publishers, Amsterdam.

A. Verma, T. Faruquie, C. Neti, S. Basu in A. Senior. 1999. Late integration in audio-visual continuous speech recognition. V: *Proceedings of Automatic Speech Recognition and Understanding (ASRU'99)*, Colorado.

M.-H. Yang in N. Ahuja. 1998. Detecting Human Faces in Color Images. V: *Proceedings of the 1998 IEEE International Conference on Image Processing (ICIP'98)*, zvezek 1, str. 127–130, Chicago.