Snimanje bilingvalne baze AI-SPEAK za multimodalno prepoznavanje govora

Tijana Nosek
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
tijana.nosek@uns.ac.rs
0000-0002-3707-0286

Siniša Suzić
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
sinisa.suzic@uns.ac.rs
0000-0002-0511-6729

Vuk Stanojev
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
vukst@uns.ac.rs
0000-0003-2517-3728

Nikša Jakovljević
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
jakovnik@uns.ac.rs
0000-0002-7283-3939

Lidija Krstanović
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
lidijakrstanovic@uns.ac.rs
0000-0001-7958-5846

Milan Sečujski
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
secujski@uns.ac.rs
0000-0002-3426-3277

***Apstrakt* – U ovom radu opisan je postupak snimanja, kao i krajnji sadržaj bilingvalne audio-vizuelne baze na srpskom i engleskom jeziku. Iako za engleski jezik postoji veći broj audio-vizuelnih baza različte veličine, ovo je prvi primer jedne takve baze na srpkom jeziku. Pored bilingvalnosti, još jedna prednost predstavljene baze ogleda se i u činjenici da su video snimci dostupni iz više različitih uglova-anfas, kao i s leve i desne strane u odnosu na glavnu kameru. Detaljno je opisan proces razvoja baze, kao i naknadna obrada i svi problem koji su se javljali.**

Ključne reči – multimodalno, audio-video, baza, govorna komunikacija.

I. UVOD

U svakodnevnom govoru, ljudi ostvaruju komunikaciju na multimodalan način, ne oslanjajući se samo na zvuk koji stiže do njih, već i na čitanje sa usana govornika, kao i na izraze njegovog lica [1]. Zbog toga se razvijaju audio-vizuelne baze podataka, koje pored govora sadrže i informacije o govornikovom licu ili samo usnama. Ovakve baze podataka omogućavaju istraživanja u različitim oblastima obrade govora, kao što su multimodalno prepoznavanje govora, multimodalna sinteza govora, konverzija video-snimaka u tekst, rekonstrukcija govora na osnovu video-snimaka i rekonstrukcija pokreta usana i ostatka lica na osnovu govora.

Iako su istraživanja o multimodalnom prepoznavanju govora sve češća, ovakvih baza podataka je trenutno veoma malo. Većina ovakvih baza podataka namenjena je isključivo za engleski jezik, uz nekoliko izuzetaka kao što su kineski i nemački jezik [2]. Obrada govora je u velikoj meri specifična za svaki jezik i da bi se postigao odgovarajući kvalitet, neophodno je razvijati kvalitetne resurse za taj jezik.

U ovom radu, biće predstavljene dostupne multimodalne govorne baze podataka, kao i detaljan način snimanja i obrade jedne takve baze, AI-SPEAK korpusa, koja pored snimaka govora 25 govornika na srpskom i engleskom jeziku, sadrži i video snimke delova njihovih lica.

II. JAVNO DOSTUPNE MULTIMODALNE BAZE

Istraživanja u oblasti prepoznavanja govora uz dodatnu informaciju u vidu video snimaka pokreta usta dovela su do formiranja različitih audio-vizuelnih korpusa snimanih u kontrolisanim uslovima. Među takvim bazama su AVLetters [3], CUAVE [4], OuluVS [5] i GRID [6], međutim, radi se o skupovima podataka sa malim brojem govornika i ograničenim rečenikom. AVLetter i CUAVE sadrže izgovore engleske abecede i 10 cifara, dok GRID i OuluVS sadrže kratke rečenice, ograničenog rečenika, iste za svakog govornika.

OuluVS2 [7] je složeniji korpus, govornici izgovaraju duže rečenice sa proširenim rečnikom i snimani su iz više uglova. AVICAR korpus [8] je po obimu sličan OuluVS2, ali se razlikuje po tome što su govornici snimani tokom vožnje automobila, iz više uglova. Da bi se premostio jaz između engleskog i ostalih jezika, razvijeni su RUSAVIC [9] korpus na ruskom jeziku i CI-AVSR [10] na kineskom jeziku, pandani AVICAR korpusu.

Ograničenje baza snimanih u kontrolisanim uslovima jeste što imaju mali broj sati govora, kao i ograničen skup reči, što nije dovoljno za potpunu obuku današnjih modela baziranih na dubokom učenju.

Da bi se razvili veći korpusi sa slobodnim rečnikom koji bolje predstavlja govor iz stvarnog sveta, razvijeni su korpusi koji uzimaju govor sa interneta sa više od 100 sati govora i velikim brojem govornika. LRW [11], LRS2-BBC [12] i LRS3-TED [13] su obimni korpusi na engleskom jeziku za audio-vizuelno prepoznavanje govora i razlikuju se pre svega po izvoru podataka i veličini. LRW-1000 [14] je obiman korpus za kineski jezik dobijen obrađivanjem TV emisija na kineskom jeziku, GLips [15] za nemački, dobijen obrađivanjem snimaka iz parlamenta. Jedna od mana ovih baza jeste što u većini slučajeva govornici nisu snimani iz više uglova.

Da bi se prevazišli nedostaci korpusa snimljenih u kontrolisanim uslovima i korpusa preuzetih sa interneta, razvijen je OLKAVS [16] korpus na korejskom jeziku sa 1107 govornika i 1150 sati govora. Govornici su snimani iz 9 različitih uglova.

III. PRIPREMA AI-SPEAK BAZE

U fazi pripreme AI-SPEAK korpusa definisano je da će ovaj korpus sadržati audio snimke govora na srpskom i engleskom jeziku od 25 odraslih govornika oba pola, kao i video snimke pokreta njihovih usana iz tri različita ugla.

Planirana količina govora po govorniku je 10 minuta. Na kraju procesa snimanja snimljeno je 35 govornika, ali finalna verzija baze, nakon procesa obrade, sadrži 30 govornika.

A. Korpus

Pripremljen je fiksan broj fonetski uravnoteženih rečenica na oba jezika, uključujući skup rečenica identičan za sve govornike i skup rečenica koje su jedinstvene za svakog govornika. Deo korpusa zajednički za sve govornike sadrži izgovore slova srpske azbuke odnosno engleske abecede, izgovore 10 cifara, 7 dana u nedelji i 13 komandnih reči („napred, nazad, levo, desno, gore, dole, potvrdi, odustani, obriši, pošalji, dalje, početak, kraj“, odnosno „forward, back, left, right, up, down, confirm, cancel, delete, send, next, home, end“), kao i 25 rečenica po jeziku. Deo korpusa jedinstven po govorniku sadrži 50 rečenica za svaki od dva jezika, i posebno je kontrolisan u pogledu fonetske pokrivenosti (oko 350 reči na srpskom i oko 400 na engleskom jeziku).

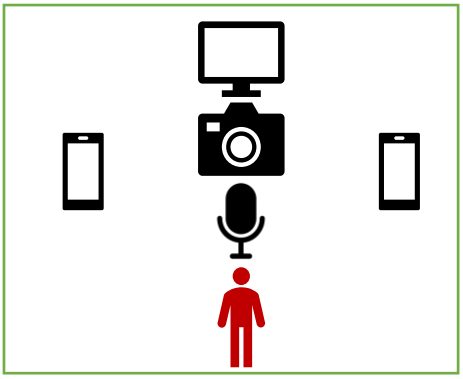
Za svakog govornika pripremljena je PPT prezentacija na kojoj svaki slajd sadrži po jednu rečenicu, a pri prelasku na sledeći slajd čuje se zvučni signal koji će se u kasnijim koracima obrade koristiti za sinhronizaciju (sinusoida frekvencije 1 kHz, trajanja 0.5 s). Govornici su instruisani da rečenicu prvo pročitaju u sebi kako ne bi pravili greške pri izgovoru te da rečenicu pročitaju normalnom brzinom, u neutralnom tonu i sa neutralnim izrazom lica, bez bilo kakvog naglašavanja u pogledu rečenične intonacije.

B. Snimanje

Snimanje je izvedeno u IAC Mini anehoičnoj komori Univerziteta u Novom Sadu. U pitanju je komora oblika kocke stranice oko 1.5m, koja je iznutra obložena modularnim akustičkim panelima koji upijaju zvuk. Po specifikaciji ova akustička izolacija smanjuje buku u odnosu na spoljašnjost komore za 50dB. Ovakve komore često se nazivaju i gluvim sobama. Ovako male gluve sobe obično su namenjene za razne eksperimente i ispitivanja u oblasti buke ili vibracija [17]. U ovom istraživanju korišćena je za snimanje govorne baze jer, pored toga što minimizuje pozadinsku buku, obezbeđena je i ponovljivost postavke opreme za snimanje i osvetljenja u uslovima kada je snimanje zbog većeg broja govornika obavljeno u dužem vremenskom periodu.

Za snimanje audio signala korišćen je Rode Podmic mikrofon. Za glavni video snimak govornikovog lica spreda korišćena je Sony VLOG ZV-1 kamera. Uz to, pomoćni audio i video snimci prikupljeni su pomoću mobilnih telefona Samsung A33 i Samsung S10, smeštenih dijagonalno ispred govornika s leve i desne strane, respektivno (Slika 1). Za osvetljenje je korišćen reflektor ugrađen u komoru, ali je prekriven odgovarajućim materijalom kako bi se formiralo difuzno svetlo.

Govornici su sedeli u anehogenoj komori leđima naslonjeni na zadnji unutrašnji zid komore, a ispred njih bio je postavljen stalak sa laptopom na kom je prikazana prezentacija sa rečenicama koje se snimaju, stalci sa kamerama i stalak za mikrofon. Celokupno snimanje



Slika 1. Šema pozicije opreme tokom snimanja u anehogenoj komori u odnosu na govornika.

po govorniku trajalo je u proseku 30-45 minuta. Govornici su instruisani da se što manje pomeraju tokom izgovora rečenice, kao i da se trude da ne menjaju položaj čitavog tela tokom celokupne sesije snimanja. U slučaju grešaka prilikom izgovora, govornici su instruisani da ponove ceo segment. Celokupna postavka u komori može se videti na slici 2. Tokom procesa snimanja, govornici su bili zatvoreni u komori, a tok snimanja praćen je van komore, i snimanje je zaustavljano ukoliko bi govornik prijavio neki problem.

Slika 2. Postavka opreme tokom snimanja u anehogenoj komori.

C. Obrada

a) Obrada audio snimaka

Kako je naknadno utvrđeno, zbog blizine električnog transformatora u susednoj prostoriji, u snimcima postoji interferencija u vidu stalno prisutnog šuma na niskim frekvencijama, pa je prvi korak obrade snimaka bila redukcija ovog šuma. Ona je urađena primenom programa *Audacity*, odnosno, odgovarajuće opcije *Noise Reduction*. U prvoj fazi bira se deo gde nema govora, odnosno gde postoji samo signal koji sadrži pomenuti šum, kako bi se izračunao spektar snage šuma. U drugoj fazi, korišćenjem infromacija o pomenutom spektru, vrši se uklanjanje šuma iz audio signala. U pojedinačnim frekvencijskim opsezima

procenjuju se snage govornog signala i šuma, a zatim se u pojedine opsege unosi odgovarajuće slabljenje u zavisnosti od toga u kojoj meri u njima dominira šum. Nakon toga vrši se vremensko ujednačavanje da bi se dobile spore promene pojačanja za svaku frekvenciju, a prati ga ujednačavanje frekvencija kako bi se postiglo da se nijedna frekvencija ne potiskuje ili pojačava izolovano.

Nakon redukcije šuma, usledio je proces anotacije koji uključuje manuelno označavanje segmenta u Audacity softveru, sa tri moguća ishoda: neupotrebljiv segment, segment sa greškama ali uglavnom ispravnim izgovorom, i segment sa ispravnim izgovorom. Segmenti su razdvojeni markerima, koje su anotatori postavljali u bilo koji trenutak tokom trajanja sinhronizacionog signala.

Tokom pregleda podataka uočene su moguće greške, kao što su: nepravilno pročitane reči, oštećeni snimci, dodati negovorni elementi kao što su zamuckivanje, tzv. lažni počeci (govornik počeo da izgovara određenu celinu, prekinuo i počeo od početka), te izgovori preklopljeni sa sinhronizacionim signalom. U zavisnosti od vrste greške, segment može biti odbačen, anotacija ispravljena, ili govornik isključen iz baze ako kod njega ima previše grešaka.

Snimci se naknadno automatski obrađuju prema podacima dobijenim manuelnom anotacijom. Prvo se vrši fino podešavanje pozicije markera na početak odnosno kraj segmenta, i to traženjem korelacije referentnog sinhronizacionog signala sa verzijom iz audio snimka. Verzija koja sadrži sinhronizacioni signal u snimku dobijena je isecanjem signala u intervalu ±1s od vremenske odrednice markera postavljenog od strane anotatora. Ovo fino podešavanje granica segmenata sprečava da se sinhronizacioni signal nađe u isečenim segmentima. Prema oznakama koje je anotator postavio, segment se ili u potpunosti zadržava, ili se odseca deo sa pogrešnim izgovorom te zadržava samo onaj sa konačnim ispravnim izgovorom, ili se u potpunosti odbacuje (Slika 3). Za delimično automatsku fonetsku i prozodijsku anotaciju (akcenti, pauze, naglasci) koristi se samo audio-snimak sa glavnog mikrofona, a dobijene transkripcije se automatski propagiraju na komplementarne snimke.

b) Obrada video snimaka

Video snimci sa svih kamera sinhronizovani su pomoću zvučnih signala snimljenih kamerama. Takođe, svi video snimci su pregledani manuelno, pri čemu su primećeni problemi sa prekomernim pomeranjem govornika tokom izgovora rečenica, kao i tehnički problemi u vidu nedostatka nekih snimaka sa pomoćnih kamera. U slučaju ozbiljnih problema, snimak se kompletno isključuje iz baze, ali se u slučaju manjih grešaka beleži napomena.

Manuelno je na video snimcima utvrđena vremenska odrednica pozicije poslednjeg sinhronizacionog signala, te su pozicije ostalih sinhronizacionih signala utvrđene odgovarajućim vremenskim pomerajem u odnosu na taj, a prema utvrđenim vremenskim odrednicama na glavnom audio-snimku. Za sinhronizaciju je odabran poslednji sinhronizacioni signal jer je on sigurno poslednji na svim kamerama i glavnom audio-signalu, dok prvi nije nužno isti jer su se neki od govornika na početku snimanja samo upoznavali sa radom prezentacije za prikaz rečenica.

Problem je takođe postojao u slučajevima kada je tokom snimanja sesije dolazilo do prekida, te ponovnog uključivanja neke od kamera iz razloga kao što su automatska zaštita od pregrevanja, što je moralo biti ručno pregledano i na odgovarajući način sinhronizovano.

Konačno, korišćenjem *MediaPipe* alata postavljene su maske preko svih video snimaka, kako bi se, radi zaštite privatnosti govornika, u AI-SPEAK korpusu našao samo donji deo lica govornika, gde se nalaze usne.

D. Najčešće greške

Oštećeni delovi audio i video snimaka su izbačeni, ali je ovoga bilo izuzetno malo (ukupno 2 rečenice). Govornici koji su se previše pomerali tokom snimanja i to tokom izgovora same rečenice (klimanje glavom, promene položaja sedenja) izbrisani su iz baze, i takvih slučajeva bilo je 3. Najveći problem bio je šta uraditi sa snimcima rečenica u kojima je pogrešna jedna reč, a govornik, suprotno uputstvima, nije ponovio celu rečenicu nego je ponovio samo tu jednu reč ili je nastavio dalje i bez pokušaja ispravke. U slučaju audio-snimka, bilo bi moguće (iako ne idealno) da se problematični deo rečenice izbaci, ali takav diskontinuitet u video snimcima apsolutno ne bi bio prihvatljiv. Kod 2 govornika ovakve greške bile su vrlo česte, te su oni izbačeni iz baze. Međutim, kod govornika kod kojih su ovakve greške bile sporadične, odgovarajući snimak bio bi izbačen u slučaju da je rečenica u delu korpusa zajedničkom za sve govornike, a u slučajevima kada se radilo o rečenici koja je namenjena samo jednom govorniku, a kada je to bilo moguće, transkript je bio modifikovan prema onome što je govornik zaista izgovorio. Utvrđeno je da postoji 2% nedostajućih vrednosti i u srpskom i u engleskom delu baze sa rečenicama zajedničkim za sve govornike (uključujući i jednu rečenicu koja je izbačena zbog oštećenog audio-snimka). U srpskom delu baze sa personalizovanim rečenicama postoji 0,3% nedostajućih vrednosti (uključujući jednu rečenicu sa oštećenim audio-snimkom), osnosno 1% rečenica sa izmenjenim transkriptom. U engleskom delu baze sa personalizovanim rečenicama ima po 1% nedostajućih rečenica i rečenica sa izmenjenim transkriptom.

IV. Rezultat

Finalni korpus se sastoji od 160 foldera po govorniku, gde svaki folder sadrži audio-snimak i tri video-snimka iste rečenice, uz transkripciju i prateće metapodatke (putanja do foldera, tekst rečenice na koju se folder odnosi, napomene

Slika 3. Korigovanje granica segmenata i prikaz mogućih scenarija (sleva nadesno): segment za brisanje, potpuno ispravan segment i segment koji sadrži i više puta pogrešno izgovorenu rečenicu, a zatim ispravno izgovorenu rečenicu.

ukoliko postoje neka odstupanja od očekivanog). Iako je snimljeno 35 govornika, u bazi je zadržano 30 govornika, 15 muških i 15 ženskih, od čega kod 3 govornika nedostaju snimci sa jedne od pomoćnih kamera. Okvirno postoji 10 minuta govora za svakog govornika za svaki od dva jezika, dakle ukupno oko 20 minuta po govorniku. Za svaki jezik je snimljeno po govorniku 80 snimaka (75 rečenica plus alfabet, cifre, dani u nedelji, pravci i komande), ali u proseku postoji po 1-2% nedostajućih vrednosti po grupi rečenica.

V. ZAKLJUČAK

AI-SPEAK baza će u okviru AI-SPEAK projekta biti javno otvorena naučnoj zajednici za istraživanje. Koliko su autori upućeni, ovo će biti prva javno dostupna bilingvalna multimodalna baza za istraživanje govora. Takođe, veoma značajan, a jedinstven deo ovakve baze čini dostupnost video-snimka usana govornika iz tri različita ugla što otvara razne mogućnosti za istraživanje. Ovaj korpus predstavlja vredan resurs za dalja istraživanja u oblasti govornih tehonologija i multimodalne analize govora.

ZAHVALNICA

Istraživanje sprovedeno uz podršku Fonda za nauku Republike Srbije, #7749, AI-SPEAK.

LITERATURA

[1] Sumby, William H., and Irwin Pollack. "Visual contribution to speech intelligibility in noise." *The journal of the acoustical society of america* 26.2 (1954): 212-215.

[2] Nemani, Praneeth, G. Sai Krishna, and Supriya Kundrapu. "Automated Speaker Independent Visual Speech Recognition: A Comprehensive Survey." *arXiv preprint arXiv:2306.08314* (2023).

[3] Matthews, Iain, et al. "Extraction of visual features for lipreading." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.2 (2002): 198-213.

[4] Patterson, Eric K., et al. "CUAVE: A new audio-visual database for multimodal human-computer interface research." *2002 IEEE International conference on acoustics, speech, and signal processing*. Vol. 2. IEEE, 2002.

[5] Zhao, Guoying, Mark Barnard, and Matti Pietikainen. "Lipreading with local spatiotemporal descriptors." *IEEE Transactions on Multimedia* 11.7 (2009): 1254-1265.

[6] Cooke, Martin, et al. "An audio-visual corpus for speech perception and automatic speech recognition." *The Journal of the Acoustical Society of America* 120.5 (2006): 2421-2424.

[7] Anina, Iryna, et al. "Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis." *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. Vol. 1. IEEE, 2015.

[8] Lee, Bowon, et al. "AVICAR: audio-visual speech corpus in a car environment." *Interspeech*. 2004.

[9] Ivanko, Denis, et al. "RUSAVIC Corpus: Russian audio-visual speech in cars." *Proceedings of the thirteenth language resources and evaluation conference*. 2022.

[10] Dai, Wenliang, et al. "Ci-avsr: A cantonese audio-visual speech dataset for in-car command recognition." *arXiv preprint arXiv:2201.03804* (2022).

[11] Chung, Joon Son, and Andrew Zisserman. "Lip reading in the wild." *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*. Springer International Publishing, 2017.

[12] Afouras, Triantafyllos, et al. "Deep audio-visual speech recognition." *IEEE transactions on pattern analysis and machine intelligence* 44.12 (2018): 8717-8727.

[13] Afouras, Triantafyllos, Joon Son Chung, and Andrew Zisserman. "LRS3-TED: a large-scale dataset for visual speech recognition." *arXiv preprint arXiv:1809.00496* (2018).

[14] Yang, Shuang, et al. "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild." *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*. IEEE, 2019.

[15] Schwiebert, Gerald, et al. "A multimodal German dataset for automatic lip reading systems and transfer learning." *arXiv preprint arXiv:2202.13403* (2022).

[16] Park, Jeongkyun, et al. "OLKAVS: an open large-scale Korean audio-visual speech dataset." *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

[17] https://www.iac-nordic.dk/downloads/lyddoede-rum/IAC%20Mini%20Anechoic%20Chambers.pdf.

Recording bilingual database AI-SPEAK for multimodal speech recognition

Tijana Nosek, Siniša Suzić, Vuk Stanojev, Nikša Jakovljević, Lidija Krstanović i Milan Sečujski

ABSTRACT

This paper describes the recording procedure as well as the final content of a bilingual audio-visual database in Serbian and English. Although there are numerous audio-visual databases of varying sizes for the English language, this is the first example of such a database in Serbian. In addition to its bilingual nature, another advantage of the presented database lies in the fact that the video recordings are available from multiple angles — frontal, as well as from the left and right sides relative to the main camera. The development process of the database is described in detail, along with the post-processing and all the issues that arose during the process.