SVEUČILIŠTE U ZAGREBU
FAKULTET ORGANIZACIJE I INFORMATIKE
VARAŽDIN

Lukas Krištić

TRAŽENJE KRIVOTVORINA U GLASOVNOM ZAPISU

DIPLOMSKI RAD

Varaždin, 2023.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Lukas Krištić

Matični broj: 0016114737

Studij: Informacijsko i programsko inženjerstvo

TRAŽENJE KRIVOTVORINA U GLASOVNOM ZAPISU

DIPLOMSKI RAD

Mentor :

Prof. dr. sc. Miroslav Bača

Varaždin, siječanj 2023.

Lukas Krištić

Izjava o izvornosti

Izjavljujem da je moj diplomski rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor potvrdio prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

Integritet i sigurnost podataka su dva vrlo važna pojma u IT sektoru. Kao što je postupak sprječavanja krivotvorenja novca u stvarnom svijetu vrlo važan tako je i sprječavanje krivotvorenja podataka u virtualnom svijetu od velike važnosti. U ovom radu ćemo se pozabaviti s krivotvorinama u glasovnom zapisu te načine pronalaska krivotvorenih glasovnih zapisa. Prvo ću se osvrnuti na povijest odnosno začetke ovog segmenta u biometriji, a zatim detaljno opisati današnje metode i načine.

Ključne riječi: Copy-Move; Splicing; Spektrogram; MFCC; Ekstrakcija svojstava; neuronske mreže; duboko učenje; krivotvorine

Sadržaj

1. Uvod

Korištenje tehnologije je danas postalo nešto neizbježno i ponekad ne uzimamo u obzir koliko vremena provodimo pred ekranima. Kako se u nekim trenutcima i sam zateknem svako malo kako uzimam mobitel u ruku i provjeravam jesam li nešto propustio od poruka, emailova ili obavijesti. Vjerujem da je dosta ljudi u sličnoj situaciji ili čak i u većem postotku koriste mobitel na dnevnoj bazi. Ovim želim dočarati kako smo u neprestanom kontaktu s uređajem koji svakodnevno prikupi masu privatnih podataka kao što su: povijest pretraživanja, poruke, audio i video zapisi, GPS lokacije i mnogo drugih. Nabrojao sam samo neke od njih ali nama su zanimljivi audio zapisi tj. glasovne poruke. Smatram da je svijest o glasovnim porukama jako mala i nismo niti svjesni kako bi se to moglo zlo upotrijebiti na sudu ili nanijeti štetu nekom poduzeću. Govorim o rekreiranju vlastitog glasa uz pomoć umjetne inteligencije, strojnog učenja, dubokog učenja čiji krajnji proizvod može biti upravo naš vlastiti glas. Ne samo kreiranje vlastitog glasa već kombiniranjem dostupnih audio zapisa ili manipulacija isto audio zapisa može nanijeti veliku štetu zbog krivotvorene poruke. [1]

Prema članku na web stranici [2], nova prevara koja se naziva (*eng. "DeepFake"*) ili (*eng. "CEO Fraud"*) postaje sve češća u poslovnom svijetu. Prevaranti koriste glasovnu imitaciju kako bi se predstavili kao direktori tvrtki i zatražili od zaposlenika da izvrše prijenos novca na račune koje kontroliraju prevaranti. Ova prevara je vrlo sofisticirana i može biti vrlo uvjerljiva, stoga je važno da tvrtke educiraju svoje zaposlenike o ovom problemu i poduzmu mjere opreza kako bi se zaštitili od ove vrste prijevare. Tako je tvrtka u Njemačkoj u jednom danu izgubila 220 000 Eura, nakon što se prevarant predstavio kao šef firme i zatražio uplatu na svoj račun. Zaposlenik je izvršio uplatu na zahtjev CEO-a, iako je u trenu posumnjao malo da je možda prevara, ali nakon što su se telefonskim putem čuli transakcija je bila obavljena. Prevarant je napravio "Deepfake" odnosno krivotvorio njegov glas kako bih došao do novaca. U kasnijem dijelu rada dotaknuti ću se teme "Deepfake" kako bih ju pobliže objasnio jer je usko povezana s temom rada, a ujedno i najveća prijetnja za audio forenziku.

Početak audio forenzike datira još od izuma fonografa Thomasa Edisona 1877. godine. Ova pionirska tehnologija omogućila je snimanje zvuka i reprodukciju putem voštanih cilindara, podižući komunikaciju na sljedeću razinu. Upotreba audio forenzike kao alata za otkrivanje skrivenih uvida u snimke zvuka datira iz 1940-ih i 1950-ih godina, kada su pravosudni organi prvi put počeli koristiti audio forenziku u kriminalističkim istragama. Navodno je najraniji oblik audio forenzike prakticirala britanska policija tijekom Drugog svjetskog rata [3]. Dok je Federalni istražni biro SAD-a (FBI) počeo je provoditi analize i poboljšanja zvuka u ranim 1960-ima. Proširujući McKeeverove principe, FBI je uspostavio postupak od 12 koraka za obradu audio zapisa.[4]

Digitalna forenzička analiza zvuka sastoji se od akvizicije, analize i procjene audio zapisa koji su dopušteni u sudu kao dokaz ili za forenzičke istrage. Digitalna multimedijalna forenzička analiza se često koristi za utvrđivanje autentičnosti i provjeru integriteta dokaza koji se podnose sudu uključujući građanske ili kaznene postupke. Glavni cilj procesa analize zvuka je postizanje jednog ili više sljedećih zadataka:

- Verifikacija integriteta ima za cilj odgovoriti na pitanje “je li upitni audio mijenjan od trenutka njegovog stvaranja ili ne?”
- Forenzičko poboljšanje zvuka ima za cilj poboljšati razumljivost govora i čujnost glasa na niskoj razini.
- Identifikacija govornika ima za cilj identificirati govornika u upitnom audio zapisu.

Razlog zbog kojeg navodim ove činjenice je način kako bih dočarao važnost ove teme, a tako ujedno istaknuo svoju motivaciju za ovaj rad, jer je vrlo zanimljivo vidjeti s kojom lakoćom se audio zapis može manipulirati. Naravno, ovo je samo kratak uvod s osnovnim informacijama i nekim zanimljivostima kako bih pokazao opasnost od audio krivotvorina te dostupnost takvih podataka. U daljnjem tekstu ćemo ući u dubinu i širinu ove teme, gdje ću objasniti koje sve metode postoje i na koji način se mogu analizirati krivotvorene audio snimke. Tako ću u drugom poglavlju govoriti o alatima, programskom jeziku Pythonu, biblioteke koje se upotrebljavaju za analizu audio snimaka koje ćemo upotrijebiti za pronalaženje krivotvorina. Poglavlje broj tri će biti srž teme te ću krenuti o vrstama audio krivotvorina, a zatim nabrojati metode koje postoje za otkrivanje krivotvorina, koje ću detaljno opisati na koji način radi. Spomenuti ću i vjerojatnost metode za pronalazak krivotvorenog zapisa i druge važne stvari. Zadnje poglavlje je praktični rad u kojem ću primijeniti navedene alate iz drugog poglavlja za prikaz analize krivotvorenog glasovnog zapisa i originalnog s jednom ili više od navedenih metoda. Na kraju ostaje još zaključak s kojim ću cijelu temu zaokružiti, iskazati što je postignuto te spojiti s vlastitim očekivanjima, poteškoćama koje sam nailazio i važnost rada.

2. Metode i tehnike rada

U ovom poglavlju će biti opisane sve metode, tehnike i alati koji su korišteni prilikom izrade ovog diplomskog rada. Za izradu i prikaz analize korišten je programski jezik Python s određenim bibliotekama koje su neophodne za analizu audio podataka, a to su Matplotlib i Librosa. Za prikaz važnosti deepfake krivotvorina korištena je web aplikacija Resemble.ai s kojom sam kreirao svoj vlastiti glas te primijenio analizu pomoću jedne od metoda za audio forenziku (*eng. "Audio Spoof"*) te usporedio ga s originalnim glasom. Na temelju kojeg bi prikazao opasnost korištenja deepfake tehnologija i njihovu težinu koja vrlo otežava pronalazak krivotvorina u glasovnom zapisu. [5]

2.1. Resemble.ai web aplikacija

Resemble AI je inovativna tehnološka tvrtka specijalizirana za generiranje prilagođenih glasova i kloniranje glasa pomoću naprednih tehnika umjetne inteligencije. Na (slici 1) je prikazan je s bijelim slovima i zelenom podlogom logo od poduzeća, a na temelju tih boja izgleda i samo sučelje web aplikacije. Osnivači Zohaiba Ahmeda i Saqiba Muhammada su napravili značajan napredak u području sinteze glasa, omogućujući korisnicima stvaranje jedinstvenih glasova visoke kvalitete za različite primjene, uključujući virtualne asistente, video igre i stvaranje sadržaja. Spomenuti ćemo njegove osnovne funkcionalnosti i temeljnu metodologiju koja pokreće njegove moćne sposobnosti generiranja glasa. Ukratko ćemo se osvrnuti i na etička pitanja koja se tiču korištenja tehnologija kloniranja glasa i pristupu tvrtke u rješavanju tih problema. Kako sam gore naveo da se poduzeće bavi razvojem prilagođenih rješenja za generiranje glasa i kloniranje glasa pomoću umjetne inteligencije (*eng. "Artificial Intelligence" AI*) i tehnika dubokog učenja. Začetak ovog poduzeća je bio 2018. godine a osnivači su Zohaib Ahmed i Saqib Muhammad, oba iskusna poduzetnika i inženjera s ljubavlju prema AI i sintezi govora. Glavna misija Resemble AI-a je omogućiti korisnicima stvaranje visokokvalitetnih jedinstvenih glasova za širok raspon primjena, od virtualnih asistenata i video igara do stvaranja sadržaja i alata za pristupačnost. [6], [7]

Slika 1: Resemble.ai logo [6]

Osnovne funkcionalnosti Resemble AI-a korisnicima pruža intuitivnu platformu za generiranje prilagođenih glasova i kloniranje glasa. Resemble AI omogućuje korisnicima stvaranje novih jedinstvenih glasova tako da prenesu malu količinu govornih podataka. Platforma koristi ove podatke za treniranje modela dubokog učenja koji generira visokokvalitetni sintetički glas koji sliči ulaznom govoru. Kloniranje glasa je proces stvaranja sintetičkog glasa koji vrlo slično oponaša glas postojećeg govornika. Platforma Resemble AI omogućuje korisnicima kloniranje glasova tako što pružaju nekoliko minuta visokokvalitetnih audio zapisa ciljanog govornika. Osim generiranja novih glasova, Resemble AI nudi alate za modificiranje postojećih glasova mijenjanjem parametara poput visine tona, brzine i emocija. Postupak generiranja (*eng. "Deep fake"*) glasa se dobiva tako da se pročita u mikrofon 10-ak izgeneriranih rečenica od aplikacije, aplikacija pokreće postupak dubokog učenja gdje nakon nekoliko minuta dobivamo model kloniranog glasa kojeg možemo koristiti za generiranje proizvoljnih rečenica. U nastavku odnosno u praktičnom dijelu će ovaj postupak biti detaljnije prikazan. Nažalost ova aplikacija nije besplatna u potpunosti, postoje tri opcije licenci a to su trial, basic i pro što je prikazano na (slika 2). Trial verzija omogućuje rekreiranje glasa ali do 15 sekundi, Basic verzija se naplaćuje $0.006 po sekundi i sadrži jedino engleski jezik. PRO verzija sadrži prijevod s jednog jezika na drugi, dodavanje emocija svom krivotvorenom glasu bez dodavanja novih audio zapisa, podržava više od 24 jezika prilikom kreiranja glasa.

	BASIC	PRO
Number of Voices	10 Included	unlimited
Team Users	unlimited	unlimited
Projects	unlimited	unlimited
Emotion Control	✓	✓
Download Audio	✓	✓
API Access	✓	✓
AI-Generated Text (Powered by GPT-3)	×	✓
Real Time Generation	×	✓
Voice Creation API	×	✓
Foreign Languages	×	✓
Streaming API	×	✓
On Premise Deployment	×	✓
Mobile Deployment	×	✓
Enterprise SLAs	×	✓

Slika 2: Licence za resemble.ai [6]

2.2. Programski jezik Python

Python je programski jezik koji je u 21. stoljeću jako rasprostranjen i korišten zbog svojih prednosti kao što su brzina izvođenja programa, jednostavnost sintakse, modularnost, podržava objektno orijentirano, imperativno i funkcionalno programiranje. Moram napomenuti da je zbog svoje jednostavnosti jako pogodan za učenje i najbolja opcija za početnike, u usporedbi s C++ ili Java jezikom koji zahtijevaju više linija koda nego što je to potrebno u Pythonu [8]. Prema web stranici TIOBE Python je u kategoriji popularnosti trenutno na prvom mjestu, prema (slika 3).Dvije godine uzastopno proglašen je 2021. i 2022. programskim jezikom svijeta po rastu populacije na godišnjoj razini popularnosti. [9]

Year	Winner
2022	C++
2021	Python
2020	Python
2019	C
2018	Python
2017	C
2016	Go
2015	Java
2014	JavaScript
2013	Transact-SQL
2012	Objective-C
2011	Objective-C
2010	Python
2009	Go
2008	C
2007	Python
2006	Ruby
2005	Java
2004	PHP
2003	C++

Slika 3: Programski jezici svijeta od 2003. do 2023. (Izvor: [9], 2023)

Interpreterski jezik Python nastao je ne tako daleke 1980. godine od strane Nizozemskog programera Guido van Rossum gdje je novo nastali jezik smatran kao nasljednik ABC jezika. Jedna od zanimljivosti za naziv Python je da potječe od humoristične serije pod nazivom Leteći cirkus Monty-ja Pythona (*eng. "Monty Python's Flying Circus"*).Na slici broj 4 Trenutna i zadnja verzija Pythona je 3.11 ali i dalje se ponekad vode rasprave za korištenjem verzije 2. jer ona je ipak imala veliku podršku od korisnika u pozadini. [5], [8]

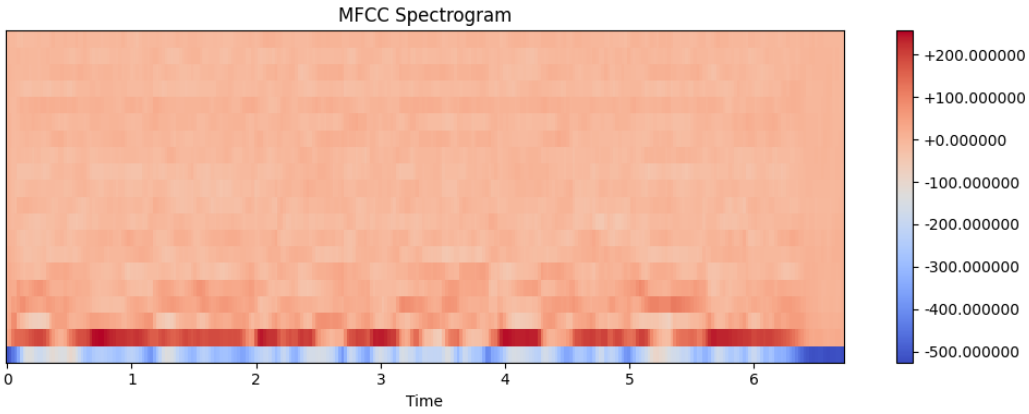
Slika 4: Python logo [10]

Mnoge karakteristike obogaćuju ovaj programski jezik, a jedna od pogodnijih karakteristika je to što je otvorenog koda (*eng. "Open source"*) što znači da ga svatko može koristiti i modificirati. Prilikom kreiranja komercijalnog proizvoda nije potrebno kupovanje licenci. Blago rečeno ako bi razvili aplikaciju u Pythonu koja bi postala komercijalna jednoga dana, sva zarada bi išla vama kao autoru aplikacije. Programeri s dugogodišnjim iskustvom u Pythonu su traženi i imaju iznad prosječne plaće u IT sektoru, dodao bih još da je sve veća potražnja za znanjem u području strojnog učenja. U prijašnjem odlomku sam već spomenuo važnost biblioteka koje krase i obogaćuju ovaj programski jezik, neke od njih služe za strojno učenje, obradu slika, obradu audio podataka. Nama će biti vrlo zanimljiv ovaj dio koji izvršava obradu audio podataka i u daljnjem tekstu ćemo reći nešto o bibliotekama Librosa i Matplotlib.[11]

2.2.1. Biblioteka Librosa

Librosa je Python paket za analizu glazbe i zvuka koji pruža širok raspon alata za izvlačenje značajki (*eng. "Feature extraction"*), uključujući spektralne značajke, poput Mel-frekvencijskih kepstralnih koeficijenata (*eng. "Mel-frequency Cepstral Coefficients"*) ili MFCC, kromatske značajke i spektralne kontrastne značajke, kao i ritamske značajke, poput tempa i praćenja ritma. MFCC je reprezentacija kratkoročnog spektra snage zvuka, temeljena na linearnoj kosinus transformaciji logaritamskog spektra snage na nelinearnoj Mel skali frekvencije slika 5, kromatske značajke se dobiju iz valnog oblika ili spektrograma. Najučestaliji audio zapisi s kojima se danas susrećemo prema članku [12] su: WAV, PCM, WMA, FLAC i MP3. Izvlačenje značajki je bitan korak u analizi glazbe i zvuka koji uključuje pretvaranje sirovih audio signala u skup smislenih značajki koje se mogu koristiti za daljnju analizu, poput klasifikacije, grupiranja ili vizualizacije. Librosa pruža jednostavno i fleksibilno sučelje za računanje širokog raspona audio značajki koje se mogu koristiti za različite primjene, poput klasifikacije žanra glazbe, prepoznavanja govora ili otkrivanja zvučnih događaja.[13], [14]

Jedna od najčešće korištenih značajki u analizi glazbe i zvuka su MFCC-ovi, koji se temelje na percepciji ljudskog sluha. MFCC-ovi se široko koriste u zadacima prepoznavanja govora i izvlačenja informacija o glazbi jer hvataju bitne karakteristike zvuka, poput visine tona, timbra i glasnoće. Druga popularna značajka u analizi glazbe su kromatske značajke, koje predstavljaju tonalitet audio signala mapiranjem na 12-tonsku kromatsku ljestvicu. Kromatske



Slika 5: MFCC spektrogram vlastitog glasa

značajke korisne su za zadatke poput prepoznavanja akorda, procjene tonaliteta i izvlačenja melodije. Spektralne kontrastne značajke su još jedan tip spektralnih značajki koje hvataju spektralni oblik audio signala računanjem razlike između njegovih vrhova i dolina. Spektralne kontrastne značajke korisne su za zadatke poput prepoznavanja instrumenata i otkrivanja zvučnih događaja.[13]

Iz priloženog vidimo da je stvorena za olakšanje složenijih zadataka obrade zvuka, nudi širok raspon funkcionalnosti koje zadovoljavaju različite aspekte analize bilo kojeg audio zapisa. U nastavku ću nabrojati kategorije po kojima se mogu klasificirati funkcije: [15] [16]

- Konverzija i učitavanje audio zapisa - Librosa dopušta korisnicima učitavanje i konverziju audio datoteka u prikaze vremenskih serija odnosno spektrograme [17]. Biblioteka podržava različite audio formate i brzine uzrokovanja, što je čini kompatibilnom s različitim audio podacima.

- Izdvajanje značajki (*eng. "Feature extraction"*) - Librosa nudi bogat skup funkcija za izdvajanje smislenih značajki iz audio signala. Neke značajke su kepstralni koeficijent mel-frekvencije (MFCC u daljnjem tekstu), spektralni kontrast i značajke boje[14]. Naveden značajke nam omogućuju prikaz visine tona, boju i harmonijski sadržaj audio signala.

- Obrada u vremenskoj (*eng. "Time-domain"*) i frekvencijskoj domeni (*eng. "frequency-domain"*): Korisnici mogu izvoditi različite operacije na audio signalima u vremenskoj i frekvencijskoj domeni koristeći Librosu. Knjižnica podržava filtriranje, pomak tonaliteta, rastezanje vremena i druge transformacije. To omogućuje korisnicima da manipuliraju audio signalima kako bi odgovarali njihovim specifičnim zahtjevima za analizu.

- Vizualizacija: Librosa nudi skup alata za vizualizaciju za istraživanje audio podataka, poput spektrograma, valnih oblika i kromagrama. Ove vizualizacije pomažu korisnicima u razumijevanju temeljne strukture i uzoraka unutar audio signala, što dovodi do dubljih analiza.

- Strojno učenje i prepoznavanje uzoraka: Značajke Librose mogu se koristiti kao ulazi za algoritme strojnog učenja, što korisnicima omogućuje izgradnju i treniranje modela

za zadatke poput klasifikacije žanra, prepoznavanja emocija i identifikacije govornika. Knjižnica se besprijekorno integrira s popularnim knjižnicama za strojno učenje poput scikit-learn i TensorFlow.[18], [19]

Opsežna funkcionalnost Librose čini je prikladnom za širok raspon primjena u obradi zvuka i analizi glazbe. Neka od značajnih primjena uključuju:

Izvlačenje informacija o glazbi (*eng. "Music information retrieval"*)(MIR): Zadaci MIR-a, poput klasifikacije žanra, prepoznavanja akorda i praćenja ritma, mogu se izvoditi pomoću mogućnosti izvlačenja značajki i strojnog učenja Librose.[16]

Obrada i analiza zvuka: Python biblioteka Librosa omogućuje korisnicima obavljanje različitih zadataka u domenama analize audio i glazbenih signala [16]. Njegove raznolike funkcionalnosti obuhvaćaju obradu govora, audio efekte i sintezu, klasifikaciju i segmentaciju audio zapisa te otkrivanje prijevara s audio zapisima.

Analiza govora: Korištenjem Librose, istraživači mogu učinkovito rješavati zadatke obrade govora poput identifikacije govornika, prepoznavanja govora i prepoznavanja emocija. Izdvajanje značajki specifičnih za govor, poput MFCC-a [14], omogućuje izgradnju robusnih modela za analizu govornog jezika.

Manipulacija i generiranje zvuka: Librosa nudi različite funkcije obrade vremenske domene i frekvencijske domene, što korisnicima omogućuje stvaranje audio efekata i sintezu novih zvukova. Ove sposobnosti imaju široku primjenu u glazbenoj produkciji, dizajnu zvuka i drugim kreativnim područjima.

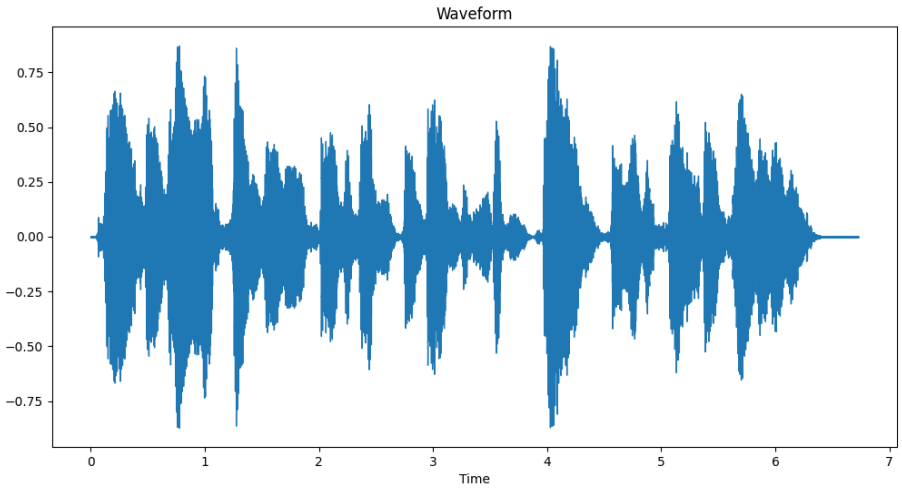
Kategorizacija i particioniranje audio zapisa: Korištenjem mogućnosti izdvajanja značajki Librose i njegove kompatibilnosti s tehnikama strojnog učenja, korisnici mogu stvarati modele za klasifikacijske i segmentacijske zadatke audio zapisa. To uključuje identificiranje zvučnih događaja i odvajanje miješanih izvora zvuka.[20]

Otkrivanje krivotvorenih audio zapisa: Istraživači mogu koristiti Librosa za razvoj metoda za identifikaciju prijevara s audio zapisima u glasovnim snimkama. Analizom značajki zvuka moguće je razlikovati između autentičnih i manipuliranih uzoraka glasa.[21]

2.2.2. Biblioteka Matplotlib

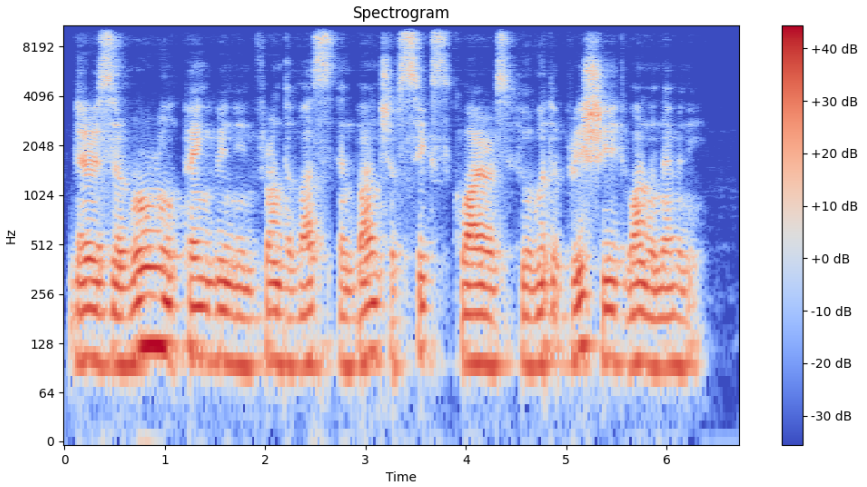
Python je postao popularan programski jezik za audio forenziku zbog svog opsežnog ekosustava znanstvenih knjižnica. Jedna takva knjižnica, Matplotlib, moćan je alat za stvaranje širokog raspona vizualizacija koje mogu pomoći forenzičkim analitičarima u razumijevanju i tumačenju audio podataka. Vrlo korisna biblioteka u području audio forenzičke analize pružajući primjere vizualizacija uz pomoć specifičnih vrsta grafova. Matplotlib je svestrana Python knjižnica razvijena za stvaranje statičkih, interaktivnih i animiranih vizualizacija. Pruža objektno orijentiran API, omogućujući korisnicima generiranje složenih vizualizacija s finom kontrolom nad izgledom i ponašanjem grafikona. Njegov sveobuhvatni raspon podržanih vrsta grafikona uključuje linijske grafikone, grafikone raspršenja, stupčaste grafikone i druge, što ga čini prikladnim za različite znanstvene i inženjerske primjene.[22]

U kontekstu audio forenzike, Matplotlib se može koristiti za generiranje vizualizacija koje otkrivaju bitne značajke i karakteristike audio signala, pomažući analitičarima da otkriju manipulirane ili krivotvorene snimke. U nastavku ću govoriti o elementima koji su neophodni za analizu zvuka kao što su audio valni oblici (*eng. "Audio waveforms"*) slika 6, spektrogrami (*eng. "Spectrograms"*) slika 7 i analiza omotnice (*eng. "Envelope analysis"*). Audio valni oblici predstavljaju amplitudu audio signala kao funkciju vremena. Pružaju uvid u cjelokupnu strukturu i dinamiku zvuka [22]. U forenzičkoj analizi vizualna inspekcija audio valnih oblika može pomoći u identifikaciji naglih promjena ili anomalija koje ukazuju na manipulaciju. Spektrogrami prikazuju frekvencijski sadržaj audio signala tijekom vremena, pri čemu se amplituda prikazuje intenzitetom boje. Ova vizualizacija ističe promjene u spektralnim karakteristikama tijekom trajanja zvuka. U audio forenzici, spektrogrami mogu otkriti neusklađenosti u frekvencijskom domenu, što sugerira moguću manipulaciju ili uređivanje. Analiza omotnice usredotočuje se na amplitudnu omotnicu audio signala, što može biti korisno za otkrivanje naglih promjena ili diskontinuiteta koji mogu proizaći iz spajanja ili drugih tehnika uređivanja [21], [23]. Matplotlib se može koristiti za prikazivanje amplitude omotnice audio signala, olakšavajući identifikaciju potencijalne manipulacije. [22]



Slika 6: Primjer grafa audio valnog oblika vlastitog krivotvorenog glasa

Iako je Matplotlib moćan alat za generiranje vizualizacija zvuka, njegove mogućnosti mogu se dodatno poboljšati integriranjem s posvećenim knjižnicama za obradu zvuka poput Librosa [16] ili Pydub. Ove biblioteke pružaju niz funkcija za izdvajanje značajki zvuka i obradu audio signala, što ih čini idealnim suputnicima Matplotlib-a u audio forenzičkim zadacima. Smatram da je nezamjenjiv alat za audio forenzičku analizu koji omogućuje forenzičkim analitičarima učinkovitiju vizualizaciju i tumačenje audio podataka. Stvaranjem vizualizacija specifičnih za audio poput valnih oblika, spektrograma i analize omotnice, analitičari mogu dobiti vrijedne uvide u karakteristike audio signala.



Slika 7: Primjer grafa spektrograma vlastitog krivotvorenog glasa

2.3. Audacity aplikacija

Audacity je besplatni softver za uređivanje digitalnog zvuka i snimanje otvorenog koda. Dostupan je za različite platforme, uključujući Windows, macOS i Linux. Audacity su prvi put objavili u svibnju 2000. Dominic Mazzoni i Roger Dannenberg sa Sveučilišta Carnegie Mellon. Softver nudi značajke za naknadnu obradu svih vrsta zvukova, uključujući podcaste. Korisnici mogu uvoziti, uređivati i kombinirati zvučne datoteke. Izvoz snimaka u mnogo različitih formata datoteka, uključujući više datoteka odjednom, također je moguć. U nastavku su nabrojane neke mogućnosti koje sve ovaj alat može obavljati i na (slika 8) je prikazan je prepoznatljiv logo aplikacije.[24]

- Snimanje: Audacity može snimati zvuk uživo putem mikrofona, miksete ili digitalizirati snimke s drugih medija.
- Uređivanje: možete rezati, kopirati, spajati ili miješati zvukove. Audacity također podržava velik broj naredbi za uređivanje putem tipkovničkih prečaca.
- Kvaliteta zvuka: Podržava 16-bitne, 24-bitne i 32-bitne (pokretni zarez) uzorke. Konverzija stopa uzorkovanja i formata je kvalitetna i profesionalna.
- Dodaci: Podrška za dodatke za efekte LADSPA, LV2, Nyquist, VST i Audio Unit. Nyquistove efekte možete jednostavno mijenjati u uređivaču teksta ili čak možete napisati vlastiti dodatak.
- Analiza: Način prikaza spektrograma za vizualizaciju i odabir frekvencija. Prikazi logaritamske/linearne skale i prozor dijagrama spektra za detaljnu analizu frekvencije.
- Formati datoteka: Podržava uvoz/izvoz brojnih formata datoteka kao što su WAV, AIFF, MP3, OGG, FLAC i drugi.

Slika 8: Logo aplikacije Audacity

3. Metode i načini pretrage krivotvorina u glasovnim zapisima

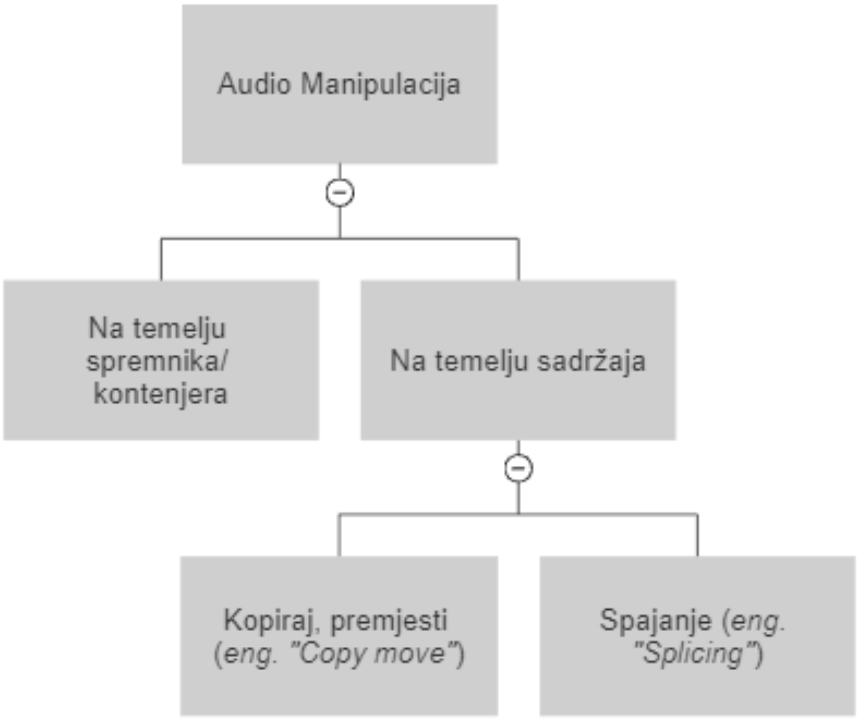
U ovom poglavlju ćemo obraditi glavnu srž ovog rada a to su vrste krivotvorina u audio ili glasovnim zapisima, zatim ćemo govoriti o metodama i načinima pretrage krivotvorina u glasovnom zapisu. Pojedine metode će biti potkrijepljene primjerom uz odabranu vrstu krivotvorine. Nadovezat ću se u ovom poglavlju i na temu deepfake krivotvorina u audio zapisima i njihove metode pronalaženja, koje su usko povezane uz ovu tematiku. Za digitalni audio zapis koji je zlonamjerno krivotvoren, bitno je razviti učinkovite metode za otkrivanje krivotvorenja digitalnog audio zapisa. Nedavno je došlo do povećanja broja softvera za digitalnu obradu zvuka koji su olakšali uređivanje, krivotvorenje i falsificiranje digitalnog zvuka. Kako je danas mnogo toga putem interneta dostupno, tako se vrlo lako mogu pronaći softveri koji pripomažu uređivanju zvuka, a neki od njih su Adobe Auditiona i Audacity. Uz malo napora i truda omogućuju običnim ljudima brisanje, umetanje, kopiranje i lijepljenje digitalnog zvuka koji je podložan krivotvorenju što dovodi do promjena u semantici zvuka. Svakodnevno poboljšanje tehnologije uređivanja zvuka zahtijeva i veću sigurnost, jer današnje manipulacije audio zapisa više nije moguće prepoznati samo pomoću ljudskog sluha, potrebni su softveri za analizu i stručno osoblje kako bi se krivotvorina razotkrila. Manipulirani digitalni audio zapisi mogu biti zloupotrijebljeni, posebno u ključnim sigurnosnim aplikacijama, na suđenju, politici ili u poslovnom svijetu, što može uzrokovati ozbiljne posljedice. Bitno je razviti učinkovite metode za otkrivanje krivotvorenja digitalnog audio zapisa za slučaj digitalnog audio zapisa koji je zlonamjerno krivotvoren[25], [26].

3.1. Vrste audio krivotvorina

Kako bih podjele metoda za detekciju krivotvorina u audio ili glasovnim zapisima bila jasnija, potrebno je prvo kategorizirati i navesti vrste audio krivotvorina. Prema Bevimaradu i društvu [27] podjela audio krivotvorina se dijeli u dvije kategorije, na temelju spremnika/kontejnera i na temelju sadržaja. Metode manipulacije audio zapisa temeljene na kontejneru su tehnike koje se koriste za manipulaciju audio datotekama mijenjanjem formata kontejnera audio datoteke. Kao primjer možemo reći da to može biti bilo kakva promjena vezana za metapodatke, opis, vremensku oznaku, format ili heksadecimalne podatke audio datoteke. Postoje dvije standardne tehnologije u aktivnim metodama detekcije koje se mogu koristiti za ovakvu vrstu krivotvorina, a to su: digitalni vodeni žig i digitalni potpis digitalnog zvuka. Naravno ove metode ćemo puno detaljnije u sljedećem poglavlju analizirati [28].

Metode audio manipuliranja temeljene na sadržaju uglavnom se usredotočuju na izmjenu izvornog audio signala kako bi se stvorio zavaravajući sadržaj, često sa zlonamjernim namjerama. Ove manipulacije mogu varirati od suptilnih uređivanja kao Kopiraj-premjesti (*eng. "Copy move"*) do naprednijih tehnika poput generiranja dubokih lažnih audio zapisa (*eng. "Deepfake"* u nastavku deepfake). Jedna uobičajena metoda je spajanje (*eng. "Splicing"*), gdje se različiti dijelovi audio zapisa preuređuju ili kombiniraju kako bi se stvorila zavaravajuća priča. Druga tehnika, promjena visine tona, mijenja visinu govornikovog glasa kako bi se sakrila nje-

gova identifikacija ili promijenila percipirana emocija. Naprednije metode, poput kloniranja ili sinteze glasa, koriste umjetnu inteligenciju i strojno učenje za generiranje zvuka koji zvuči realistično iz teksta ili oponašaju glas ciljanog govornika. Metode za audio manipulacije temeljene na sadržaju predstavljaju značajne izazove za digitalnu forenziku jer postaju sve sofisticiranije i teže ih je otkriti, što izaziva zabrinutost u vezi autentičnosti i vjerodostojnosti audio sadržaja u digitalnom dobu. Kako bi se bolje dočarala ova podjela na (slici 9) možemo pogledati opisanu klasifikaciju. U daljnjem tekstu ćemo opisati neke od metoda koje spadaju pod kategoriju audio manipulacije na temelju sadržaja i dodao bih još jednu granu uz postojeće dvije, a to je grana za generiranje audiofake-ova . Koja se dijeli na audio zapise Generirane bez umjetne inteligencije (*eng. "Non AI Generated"*) i audio zapisi generirani pomoću umjetne inteligencije ili deepfake-ovi ((*eng. "AI Generated (Deepfake) "*)) [29].



Slika 9: Klasifikacija audio krivotvorina (Izvor: Bevinamarad i Shirldonkar, 2020)

3.1.1. Manipulacija audio zapisa na temelju sadržaja

Metode manipulacije zvuka temeljene na sadržaju uključuju izravno manipuliranje audio signalom kako bi se stvorio obmanjujući sadržaj ili promijenila izvorna poruka. Te tehnike mogu varirati od jednostavnog uređivanja do naprednijih procesa potaknutih umjetnom inteligencijom. U nastavku su nabrojane tehnike za ovu kategoriju:[27]

- Kopiraj-premjesti (*eng. "Copy move"*)

- Spajanje (*eng. "Splicing"*)

3.1.1.1. Kopiraj-premjesti (*eng. "Copy move"*)

Tehnika manipulacije zvukom kopiraj-premjesti sve je češća metoda u digitalnoj forenzici zvuka koja uključuje izravnu izmjenu audio sadržaja kroz dupliranje određenih segmenata i njihovo naknadno umetanje na različitim mjestima unutar iste datoteke. Ova tehnika služi različitim svrhama, poput prikrivanja nedostataka, stvaranja besprijekornih petlji ili repliciranja određenih zvukova. Široka dostupnost naprednog softvera za uređivanje zvuka pridonijela je rastućim izazovima u otkrivanju i analiziranju kopirano-premještenog sadržaja. Ljudskim preslušavanjem ovakvog tipa krivotvorenog audio zapisa se vrlo teško može identificirati manipulacija jer ovakve snimke govora mogu biti jako duge. Ako bih uzeli rečenicu "Jučer je bio lijep dan, a danas nije padala kiša." možemo uzeti riječ "nije" i postaviti ju umjesto riječi "je" tada bi rečenica imala potpuno drugo značenje "Jučer nije bio lijep dan, a danas nije padala kiša.", u ovom primjeru vidimo da duplicirani govorni segmenti mogu biti sažeti poput samo jedne riječi. Ako bi uzeli da je glasovna poruka duljine 30-60 minuta, bilo teško uočiti ovakvu manipulaciju, a dodatno se mogu koristiti procesi naknadne obrade za brisanje tragova krivotvorenja. Znanstvenici u području digitalne forenzike zvuka razvijaju inovativne algoritme za suprotstavljanje ovoj tehnici pregledavanjem suptilnih neusklađenosti u dupliranim segmentima, uključujući varijacije u pozadinskom šumu, spektralnim značajkama ili akustičkim svojstvima. Ove metode detekcije često uključuju strojno učenje i umjetnu inteligenciju kako bi se poboljšala njihova točnost i prilagodljivost u identificiranju manipulacija zvukom kopiraj-premjesti. Neprestani napredak tehnika digitalnog uređivanja zvuka zahtijeva kontinuirano istraživanje i inovacije kako bi se održala autentičnost i pouzdanost sadržaja u brzo se razvijajućem digitalnom okruženju.[27], [30]–[32]

3.1.1.2. Spajanje (*eng. "Splicing"*)

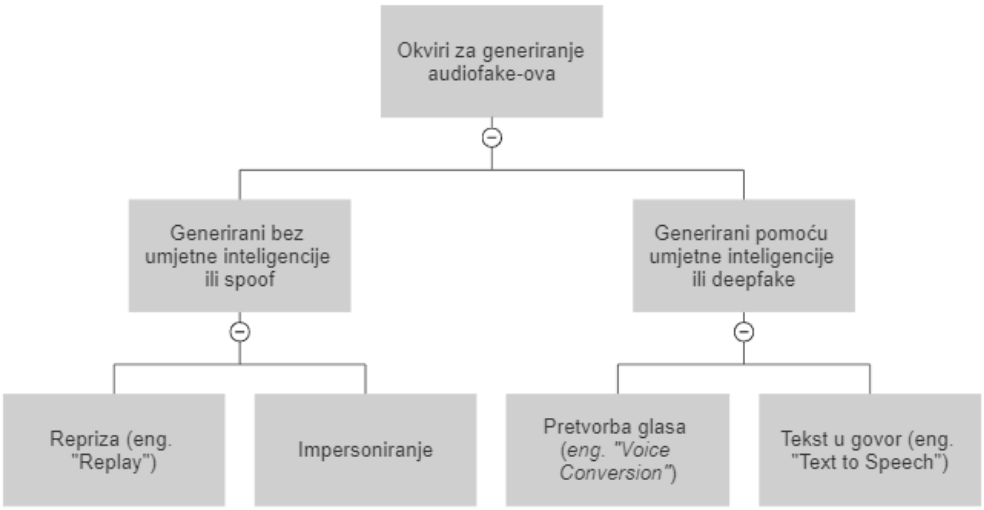
Tehnika manipulacije audio spajanjem (*eng. "Splicing"*) složena je metoda u području digitalne forenzike zvuka koja uključuje proces segmentiranja i preuređivanja audio zapisa kako bi se stvorile izmijenjene naracije ili konteksti. Ekstrahiranjem specifičnih dijelova audio signala i ponovnim sastavljanjem u drugačijem redoslijedu, ova tehnika može učinkovito izmijeniti izvorni sadržaj, često s namjerom zavaravanja slušatelja ili prikazivanja netočnih informacija. Sve veća dostupnost naprednih softvera i alata za uređivanje učinila je spajanje dostupnijim i izazovnijim za otkrivanje. Reverberacijom i dodavanjem buke uključuju namjernu degradaciju kvalitete zvuka, što otežava forenzičkim analitičarima provjeru autentičnosti snimke ili prepoznavanje neovlaštenog mijenjanja. Dodavanjem pozadinske buke ili manipuliranjem karakteristikama odjeka, počinitelj može stvoriti vjerojatnu mogućnost poricanja bilo kakvih nedosljednosti ili izmjena u zvuku. Ovi pristupi obično se oslanjaju na analizu spektralnih i temporalnih značajki, kao i na ispitivanje neusklađenosti u pozadinskom šumu, odjeku i drugim akustičkim svojstvima. Budući da se tehnike spajanja nastavljaju razvijati, kontinuirana istraživanja i inovacije ključni su za održavanje integriteta i autentičnosti audio sadržaja u digitalno doba obilježeno brzom diseminacijom informacija. U ovakvim krivotvorinama ponekad ključnu ulogu imaju artefakti

u mikrofonima nakon snimanja ili ambijent u kojem se trenutno nalazi mikrofon ili pozadinska buka, jer na temelju toga postoji mogućnost da će se krivotvorina identificirati. Ova tehnika je jedna od najčešće korištenih za manipulaciju audio zapisa. [27], [33]–[35]

3.1.2. Audiofake krivotvorine generirane pomoću okvira

Generiranje audiofake okvira pojavilo se kao značajno područje zabrinutosti u digitalnoj forenzici audio zvuka zbog njihovog potencijala za proizvodnju lažnog i manipuliranog sadržaja. Ovi okviri koriste napredne tehnologije, poput umjetne inteligencije, strojnog učenja i dubokog učenja, kako bi sintetizirali visoko realistične i uvjerljive krivotvorene audio zapise. Pod ovu kategoriju spadaju tehnike poput kloniranja glasa, sinteze teksta u govor i generiranja deepfake audio zapisa. Nabrojane metode postaju sve sofisticiranije, omogućujući stvaranje sadržaja koji oponaša nijanse ljudskog govora i teško se razlikuje od autentičnih zapisa. Znanstvenici iz područja digitalne forenzike zvuka aktivno razvijaju metode za otkrivanje i suzbijanje ovih lažnih zapisa analizirajući neusklađenosti u generiranom zvuku, poput spektralnih značajki, vremenskih obrazaca i akustičkih svojstava. Osim toga, istražuju primjenu modela strojnog učenja, poput neuronskih mreža i algoritama dubokog učenja, kako bi se poboljšala točnost i učinkovitost otkrivanja lažnih audio zapisa. Kako se razvoj i dostupnost okvira za generiranje lažnih audio zapisa nastavlja napredovati, važno je da znanstvena zajednica prioritetno istražuje i inovira kako bi se održao integritet i autentičnost audio sadržaja u sve složenijem digitalnom svijetu. U nastavku je prikazana (slici 10) i napisana podjela okvira za generiranje audiofake-ova: [29]

- Generirani bez umjetne inteligencije ili spoof
  - Repriza (*eng. "Replay"*)
  - Impersoniranje
- Generirani pomoću umjetne inteligencije ili deepfake
  - Pretvorba glasa (*eng. "Voice Conversion"*)
  - Tekst u govor (*eng. "Text to Speech"*)



Slika 10: Klasifikacija audiofake krivotvorina (Izvor: Khanjani, Watson i Janeja, 2022)

3.1.2.1. Repriza (*eng. "Replay"*)

Metoda napada lažiranja zvuka pod nazivom repriza je sofisticirana prijetnja kibernetičkoj sigurnosti koja cilja glasovne komunikacijske sustave i protokole provjere autentičnosti. U ovoj vrsti napada zlonamjerni akter presreće, snima i manipulira izvornim audio podacima, kao što su glasovne naredbe ili biometrijske informacije, s namjerom da prevari ciljni sustav ili korisnike. Ponavljanjem izmijenjenih audio podataka napadač može dobiti neovlašteni pristup, lažno se predstavljati kao legitiman korisnik ili izvršiti zlonamjerne radnje na ugroženom sustavu. Audio lažni napadi posebno su podmukli jer iskorištavaju ljudsko oslanjanje na slušne znakove za povjerenje i identifikaciju. Kako bi se obranile od tih napada, organizacije moraju usvojiti višeslojni sigurnosni pristup koji uključuje naprednu analizu zvuka, otkrivanje anomalija i kontinuirane metode provjere autentičnosti korisnika kako bi učinkovito ublažile rizike povezane s ponavljanjem napada lažiranjem zvuka. [29], [36]–[38].

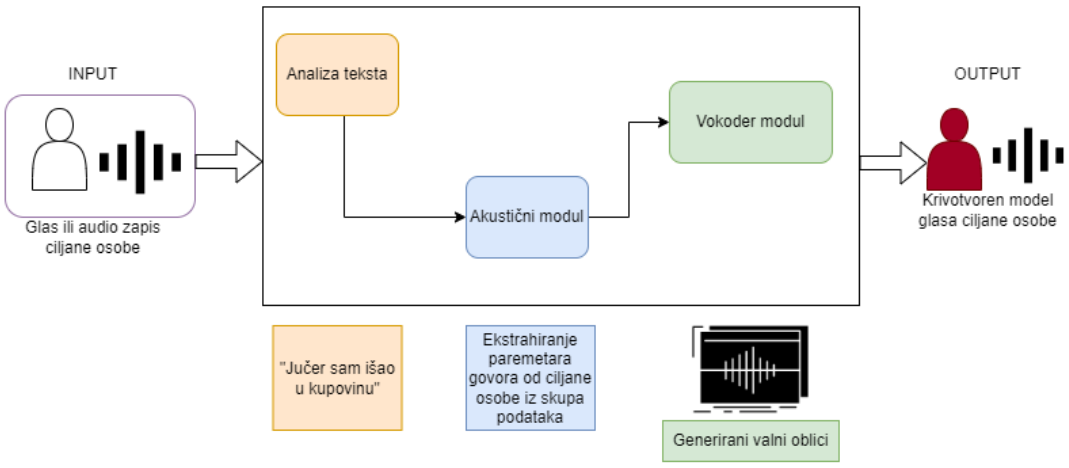
3.1.2.2. Impersoniranje

U napadima govorom ili mimikom, govorimo o vrlo talentiranim osobama koje su sposobne namjerno modificirati svoj govor da zvuči kao govor od ciljane osobe. Impersoniranje zahtijeva vrlo vješte pojedince koji su u stanju kopirati leksičko, prozodijsko i idiosinkratično ponašanje ciljnih govornika. Ovakav pristup predstavlja potencijalnu točku ranjivosti koja se tiče sustava za prepoznavanje govornika. Akterima koji su sposobni koristiti pristup impersoniranja, nije potreban nikakav softver ili alati prilikom izvođenja nedozvoljenog čina poput biometrijske autentifikacije. Potrebno je mnogo vremena kako bi netko u potpunosti naučio leksičke i prozodijske elemente drugog govornika, zbog toga se ova metoda predstavljanjem ne smatra se uobičajenom prijetnjom sustavima verifikacije govornika.[37], [38]

3.1.2.3. Tekst u govor (*eng. "Text to Speech"*)

Audio spoofing napad sinteze govora (*eng. "Speech synthesis"*) odnosi se na vrstu kibernetičke prijetnje u kojoj zlonamjerni akter koristi tehnologije sinteze govora koje se pokreću umjetnom inteligencijom, poput pretvaranja teksta u govor (*eng. "Text to Speech"*) ili kloniranja glasa, kako bi oponašao glas određene osobe. Ova napredna metoda lažiranja zvuka predstavlja značajan rizik, posebno u sustavima koji se oslanjaju na glasovnu autentifikaciju. Generiranjem visoko realističnih glasovnih izlaza, napadači mogu prevariti sigurnosne sustave ili pojedince, dobiti neovlašteni pristup osjetljivim informacijama ili provodeći lažne aktivnosti pod krinkom lažno predstavljene osobe. Takvi napadi iskorištavaju povjerenje u glas kao oblik identiteta, čineći tradicionalne sigurnosne mjere neučinkovitima. Tekst u govor metoda prima tekst kao ulazni parametar od kojeg generira glas pomoću dubokog učenja tj. kreira model kojem možemo poslati bilo kakav tekst koji će biti snimljen s našim glasom. Postoje i poneki nedostaci ove metode, a to je da ponekad ne može prepoznati naglasak kada dvije riječi imaju različito značenje, a isto se pišu. Svijest o interpunkcijskim znakovima u tekstu su također zanemareni u ovoj metodi.[29], [36]–[38]

Za kreiranje sintetičkog glasa potrebna su 3 modula: modul analiziranja teksta, akustični modul i vokoder modul. Modul za analiziranje teksta obrađuje dobiveni tekst te ga konvertira u lingvističke karakteristike. Akustični modul radi ekstrahiranje parametara ciljanog glasa iz skupa podataka na temelju lingvističkih značajki iz prethodnog koraka. Posljednji model je vokoder koji na temelju dubokog učenja kreira valne oblike govora prema parametrima akustičkih značajki iz drugog koraka. Cijeli postupak je prikazan na (slici 11). [36]



Slika 11: Tekst u govor proces (Izvor: Almutairi i Elgibreen, 2022)

3.1.2.4. Pretvorba glasa (*eng. "Voice Conversion"*)

Audio spoofing napad pretvorba glasa poseban je oblik kibernetičke prijetnje gdje napadač manipulira svojim glasom kako bi oponašao glas druge osobe. To se radi pomoću naprednih digitalnih tehnologija i softvera. Cilj je prevariti sustave koji se oslanjaju na prepoznavanje

glasa za autentifikaciju korisnika, kao što su glasovno aktivirani digitalni asistenti, glasovno upravljani sigurnosni sustavi ili telekomunikacijske usluge. Napadač može uvjerljivim oponašanjem glasa zaobići sigurnosne protokole, dobiti neovlašteni pristup ili izvoditi radnje pod maskom lažne osobe. Pretvorba glasa obično se postiže naprednim tehnikama strojnog učenja, često korištenjem algoritama dubokog učenja za stvaranje uvjerljive imitacije jedinstvenih vokalnih karakteristika cilja. Većina pristupa pretvorbi glasa zahtijeva dva glasa, od kojih je jedan izvorni a drugi od ciljane osobe te zahtijevaju se identični izgovori. Nakon faze treniranja, primjenjuju se transformacijske funkcije koje pretvaraju akustične parametre izvornog govornika onima ciljnog koje se naziva "paralelna pretvorba glasa". Neki nedostaci glasovne pretvorbe uključuju fonetske probleme, prozodiju, kvalitetu, sličnost, prekomjerno opremanje i prijetnje sustavima. Kako bi se ublažila ova vrsta prijetnje, potrebni su dodatni slojevi sigurnosti kao što su autentifikacija s više faktora, bihevioralna biometrija i kontinuirane metode provjere. [29], [36]–[38]

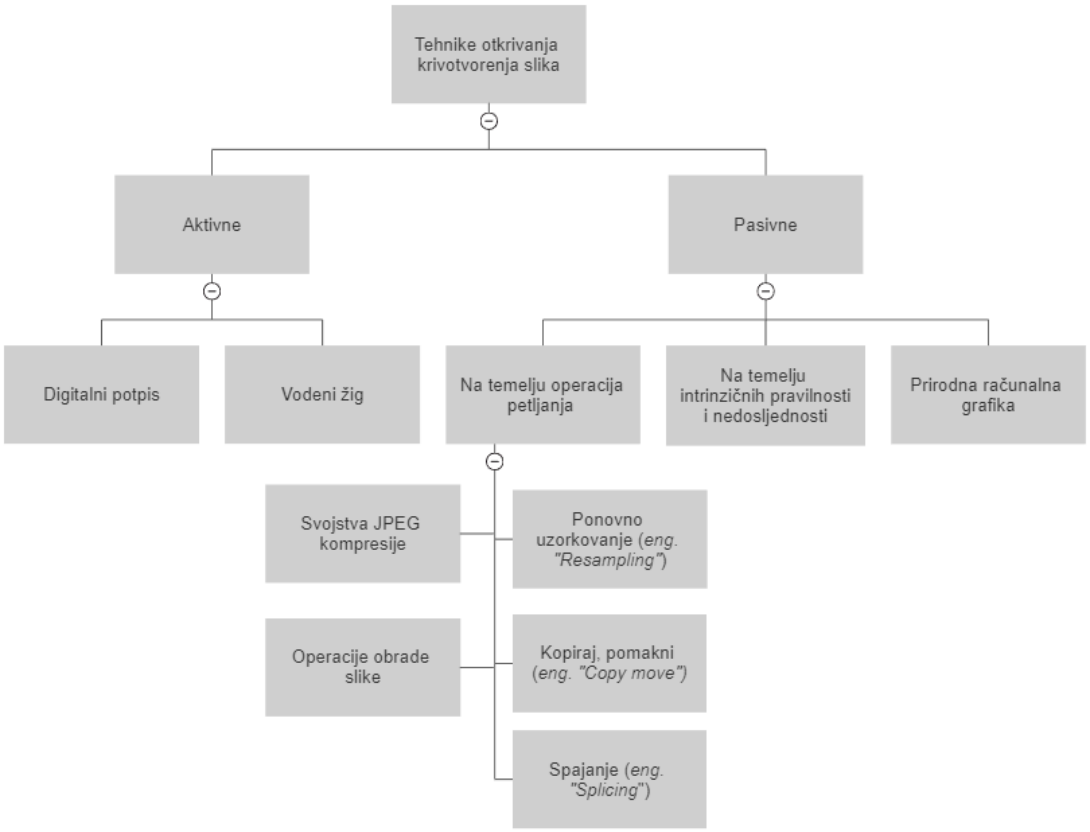
3.2. Podjela metoda detekcije krivotvorina u audio zapisima

Postoji nekoliko podjela s određenim metodama koje se upotrebljavaju kod detekcije krivotvorina audio zapisa, prema Zengu, Gupti, Wangu i društvu [28], [39], [40] podjela se klasificira na aktivne i pasivne metode. Pasivne i aktivne metode detekcije krivotvorina u audio zapisima odnose se na različite pristupe kojima se pokušava identificirati manipulacija ili krivotvorenje audio materijala.

Aktivne metode za pronalaženje krivotvorina u audio zapisima uključuju dodavanje dodatnih informacija ili značajki u zapis kako bi se olakšalo pronalaženje manipuliranih audio zapisa. Takve metode pomažu osiguravanju autentičnosti, integriteta i neporecivosti audio materijala, što je vrlo korisno kod suđenja ili policijskih istraga. Aktivne metode su: digitalni vodeni žigovi, kriptografski/digitalni potpisi, audio stenografija i korisnička provjera. [28], [39], [40]

Pasivne metode detekcije krivotvorina u audio zapisima oslanjaju se na analizu inherentnih svojstava audio materijala kako bi se otkrile nepravilnosti ili znakovi manipulacije. Ove metode ne zahtijevaju dodavanje dodatnih informacija ili značajki u audio zapis. Pasivne metode su: analiza valnog oblika, spektrogram analiza, analiza reverberacije, analiza prozodijskih svojstava, analiza formanata, detekcija na temelju buke, strojno učenje i umjetna inteligencija. [28], [39], [40]

U daljnjem radu će svaka metoda ukratko objašnjena i na kojim principima radi, pojedine metode će biti potkrijepljene primjerima. Podjela na aktivne i pasivne metode vrlo slična onoj za detekciju kod krivotvorina slika, koja se također dijeli na pasivne i aktivne metode, ali to je zato što je pronalaženje krivotvorina u audio zapisima poprilično mlada grana i pojedine metodologije, tehnike su preuzete od tamo [41]. Na (slici 12) je prikazana klasifikacija metoda za pronalaženje krivotvorina u digitalnim slikama.



Slika 12: Klasifikacija metoda za pronalaženje krivotvorina u digitalnim slikama (Izvor: Birajdar i Mankar, 2013)

3.2.1. Digitalni vodeni žigovi (*eng. "Digital Watermarking"*)

Digitalni audio vodeni žig je tehnika koja se koristi za umetanje skrivenih podataka u audio datoteku, na način koji je obično neprimjetan slušatelju. Ugrađeni podaci, poznati kao vodeni žig, mogu se koristiti u razne svrhe kao što su zaštita autorskih prava, provjera autentičnosti sadržaja i praćenje podataka. Izazov u digitalnom audio vodenom žigu je dizajn vodenog žiga koji je robustan, što znači da može preživjeti različite operacije obrade signala (poput kompresije, filtriranja ili pretvorbe), a opet neprimjetan kako ne bi degradirao kvalitetu izvornog zvuka. Osim toga, također bi trebao biti siguran od namjernih napada usmjerenih na uklanjanje ili promjenu vodenog žiga. Tehnike za digitalni audio vodeni žig često uključuju manipulaciju audio signala u frekvencijskoj domeni ili korištenje psiho akustičkih modela za skrivanje vodenog žiga u dijelovima signala gdje je najmanje vjerojatno da će ga ljudsko uho percipirati. Zbog sofisticiranosti koja je potrebna u ovim tehnikama, digitalni audio vodeni žig je aktivno područje istraživanja u poljima obrade signala i informacijske sigurnosti.

Tako su Muroi i društvo ponudili rješenje za digitalne vodene žigove tako što su izvadili "otiske prstiju" iz audio snimke i ugradili ih u audio snimke kao vodeni žigovi za otkrivanje neovlaštenog manipuliranja. Ako decoder ne pronađe vodeni žig koji bi trebao biti ugrađen u audio zapisu tada će ga svrstati kao krivotvorinu. Za ovu metodu korišten je Wiener-ov filter koji bi trebao pružiti visoku kvalitetu potiskivanja nestacionarne buke kako bi s preciznošću mogao izvući digitalni žig. Prilikom provjere integriteta audio zapisa otisak koji je ugrađen kao vodeni žig se

izdvaja i dekodira pomoću tajnog ključa. Iz reproduciranog audio zapisa koji sadrži vodeni žig, ekstrahiramo također otisak. Tada uspoređujemo izračun udaljenosti između dva otiska, te na temelju zadanog praga odlučujemo je li audio krivotvoren ili ne. Postupak ugrađivanja digitalnog žiga je takav da audio zapis podijelimo na okvire i zatim iz svakog okvira izdvojimo otisak. Na temelju (*eng. "Line-Spectral-Pairs "* LSP) su otisci izrađeni koji je posebno koristan za kodiranje, kompresiju i sintezu govora zbog svoje učinkovite reprezentacije, glatke interpolacije i stabilnosti. LSP-ovi pružaju kompaktan i robustan prikaz prijenosne funkcije vokalnog trakta, koja karakterizira rezonantne frekvencije vokalnog trakta tijekom proizvodnje govora. Metoda korištena za digitalni žig je (*eng. "Direct Sequence Spread Spectrum"* DSSS)[42]. DSSS je metoda koja se koristi u digitalnoj obradi signala koja širi signal preko šireg frekvencijskog pojasa od izvornog signala, poboljšavajući njegovu otpornost na smetnje i prisluškivanje. Kada se primijeni na digitalni audio vodeni žig, DSSS može pružiti robustan i neprimjetan vodeni žig.[43]. Ova metoda se pokazala vrlo preciznom kod pronalaženja manipulacija u audio zapisima čak 99% točnost, unatoč dodavanju buke, oponašanja jeke, 150 vrsta audio zapisa i podešavanja na 50 dB-a.

3.2.2. Digitalni potpisi ili otisci

Ova metoda uključuje korištenje kriptografskih algoritama za generiranje digitalnog potpisa koji se dodaje audio zapisu. Digitalni potpis garantira autentičnost, cjelovitost i neporecivost zapisa. Ako se zapis mijenja, digitalni potpis više neće biti valjan, što upućuje na manipulaciju.

Sustav za otkrivanje manipulacije zvuka temeljen na hash-u radi stvaranjem jedinstvenog digitalnog otiska prsta ili hash-a audio datoteke, koji se zatim može koristiti za identifikaciju kopija ili izvedenica te datoteke. Proces kreiranja otiska ili digitalnog popisa kreće s ekstrakcijom značajki (*eng. "Feature Extraction"*). Sustav prvo izdvaja značajke iz audio datoteke koje su važne za ljudski sluh i otporne na uobičajene vrste izobličenja poput kompresije, ekvilizacije ili vremenskog rastezanja. To mogu biti spektralne značajke poput veličine određenih frekvencijskih pojasa, vremenske značajke poput omotnice amplitude ili druge vrste značajki ovisno o specifičnom sustavu. Drugi korak je raspršivanje (*eng. "Hashing"* odnosno pretvaranje značajki u hash vrijednosti. hash vrijednost je kratak niz bajtova fiksne duljine koji je jedinstven za audio datoteku. Funkcija raspršivanja osmišljena je na takav način da će čak i male promjene u audio datoteci proizvesti drastično drugačiji hash. Međutim, hash bi trebao ostati isti za iskrivljenja protiv kojih bi sustav trebao biti otporan. Nakon hashiranja potrebna je pohrana u bazu podataka, u ovom koraku je vrlo važno da hashovi budu zajedno s metapodacima (duljina, naziv i sl.) o izvornim audio datotekama pohranjeni.Postupak detekcije kreće tako da sustav izdvaja značajke iz audio zapisa, stvara hash, a zatim pretražuje bazu podataka za odgovarajući hash. Ako se pronađe podudaranje, sustav zaključuje da je nova datoteka kopija originala ili izvedena od originala. Posljednji korak je naknadna obrada gdje se obavlja rješavanje hash kolizija gdje različite audio datoteke proizvode isti hash ili za pružanje više informacija o tome kako se nova datoteka odnosi na izvornu. Ovakav princip rada se koristi kao kod aplikacije Shazam. [44], [45]

Za kreiranje hash vrijednosti korišten je Chromaprint, program otvorenog izvornog koda za audio otiske, generira hash vrijednost za otkrivanje manipulacije u audio zapisu. Metoda je testirana s dodatnim poteškoćama kao što je povećanje buke 0-20 dB, tonalitet je mijenjan u rasponu od 99% do 96% od izvornog. Na temelju ovih prepreka metoda detekcije HADES je uspjela doseći 75% vjerojatnost pronalaženja krivotvorina. Iako je znatno lošije pronalazila kada se mijenjao tonalitet u audio zapisu uspoređujući s ostalim elementima. [45]

3.2.3. Korisnička provjera

Aktivna metoda koja uključuje angažiranje korisnika u procesu verifikacije audio zapisa. To može uključivati metode poput slušanja i uspoređivanja audio zapisa s poznatim autentičnim uzorcima ili upotrebu tehnikâ poznavanja tajnih informacija koje su poznate samo autoru i korisniku. Ova metoda zahtijeva stručno znanje audio analitičara koji može prepoznati krivotvorine ako se radi o spajanju više glasovnih zapisa ili nekoj drugoj navedenoj krivotvorini.

3.2.4. Spektrogram analiza

Spektrogram analiza je vizualni prikaz frekvencijskog spektra zvuka koji može otkriti neuobičajene promjene ili diskontinuitete koji upućuju na manipulaciju. Spektrogram je vizualni prikaz frekvencijskog sadržaja signala tijekom vremena. Analizom spektrograma glasovne snimke možemo identificirati sve anomalije ili nedosljednosti koje mogu ukazivati na neovlašteno mijenjanje. Na primjer, možemo tražiti promjene u spektralnoj gustoći, promjene u distribuciji frekvencija ili diskontinuitete u spektrogramu.

Tehniku spektrogram analize su istražili Ulutas i dr. koji su kreirali metodu detekcije copy-move krivotvorina pomoću (*eng. "Gray-level Co-Occurrence Matrix"* GLCM) matrice i Mel spektrograma. Prvi korak u ovoj metodi je kreiranje Mel spektrogram slika za svaki krivotvoreni audio zapis tako da koristimo (*eng. "Short-Time Fast Fourier (STFT)"*) transformaciju iz formule 1. wn predstavlja Hamming window funkciju, xn je originalni signal glasovnog zapisa, a f predstavlja frekvencijski razmak za $k = 1, 2, ...N/2 + 1$, i N je broj uzoraka u jednom okviru. [46]

$$S(f,t) = \sum_{n=0}^{N-1} w_n x_t(n) \exp(-j2\pi(f/f_s)n), f = kf_s/N \tag{1}$$

GLCM metoda je statistička metoda koja se koristi za ispitivanje teksture slike. Izračunava koliko se često različite kombinacije vrijednosti svjetline piksela (razine sive) pojavljuju na slici ili regiji slike. Računaju se matrice za kontrast, korelaciju, energiju i homogenost. Metoda uključuje dijeljenje slike spektrograma u blokove koji se preklapaju, iz kojih se značajke izdvajaju (*eng. "Feature Extraction"*) korištenjem matrice istodobnog pojavljivanja razine sive GLCM za svaki kanal boje (crvena, zelena, plava). Ove značajke, uključujući kontrast, korelaciju, energiju i homogenost, tvore matricu vektora značajki koja se zatim sortira leksikografski, omogućujući učinkovitu usporedbu sličnih vektora i ubrzavajući proces otkrivanja krivotvorina. Metoda traži slične blokove unutar matrice koji odgovaraju dupliciranim regijama, koristeći euk-

lidsku udaljenost za mjerenje sličnosti između vektora obilježja. Nakon identificiranja potencijalnog krivotvorenja, izračunava vektore pomaka između sličnih blokova i sprema ih. Ako broj krivotvorenih parova blokova s istim vektorima pomaka premaši unaprijed definirani prag, ti se blokovi označavaju kao krivotvoreni. Za pročišćavanje rezultata, metoda koristi algoritam označavanja povezane komponente za identifikaciju i zadržavanje dvije najveće regije premještene kopijom dok uklanja manje komponente, označavajući odgovarajuća krivotvorena područja u audio ulazu.[46]

3.2.5. Analiza reverberacije

Akustična reverberacija je odraz zvuka koji do slušatelja stiže sa zakašnjenjem nakon izravnog zvuka. Jednostavno rečeno, to je jeka koju čujete kada se zvuk odbija od zidova sobe. Karakteristika je prostora u kojem je zvuk snimljen, te stoga može dati ključnu informaciju o autentičnosti glasovnog zapisa.

Praktični primjer za ovu metodu su nam demonstrirali Malik i Farid te njihov drugi scenarij je izgledao tako da su generirali audio snimke s različitim količinama reverberacije ili vremena odjeka $\tau = 0,3$ ili $\tau = 0,6$ sekundi) korištenjem specifičnog modela. Svaka je snimka bila duga 9 sekundi i bila je oštećena dodatnim bijelim šumom s omjerom signal-šum od 35 dB. Hibridne snimke nastale su spajanjem dviju snimki, svaka s različitim vremenima odjeka, u jednu snimku. Spajanje je bilo nečujno jer nije bilo zvučnog spoja gdje su snimke kombinirane. Ručno je odabrano 8 pozicija vremena odjeka na temelju slabljenja govora do razine buke. U oba slučaja, prva i druga polovica zvuka imale su različita vremena odjeka. Za prvu hibridnu snimku, prva polovica je imala vrijeme odjeka od 0,3 sekunde, a druga polovica je imala vrijeme od 0,6 sekundi. U drugom hibridnom snimanju ta su vremena bila obrnuta. Srednji procijenjeni parametar opadanja τ pokazao je značajnu razliku između prve i druge polovice svakog hibridnog zapisa. Ta bi razlika mogla poslužiti kao dokaz manipulacije audio snimkama. U trećem eksperimentu, ljudski govor snimljen je u četiri različita okruženja (vani, mali ured, veliki ured, stubište) s istim govornikom koji je čitao odlomak iz Priče o dva grada Charlesa Dickensa. Nakon primjene filtra za poboljšanje govora za smanjenje pozadinske buke, odjek je procijenjen s četrnaest mjesta u svakoj snimci. Rezultirajuće srednje procjene parametra slabljenja τ pokazale su značajne razlike u različitim okruženjima, što sugerira da tekući prosjeci za kratku duljinu zvuka mogu otkriti razlike u odjeku, što potencijalno ukazuje na okruženje snimanja. [47]

3.2.6. Detekcija na temelju pozadinske buke

Metoda detekcije na temelju pozadinske buke može biti korištena za otkrivanje krivotvorina koje su kreirane spajanjem (*eng. "Splicing"*). Analizom pozadinske buke u cijelom glasovnom zapisu može se otkriti nepravilnosti koje upućuju na manipulaciju. U nastavku su definirane funkcije koje se koriste u navedenoj metodi, a zatim je objašnjen proces detekcije.

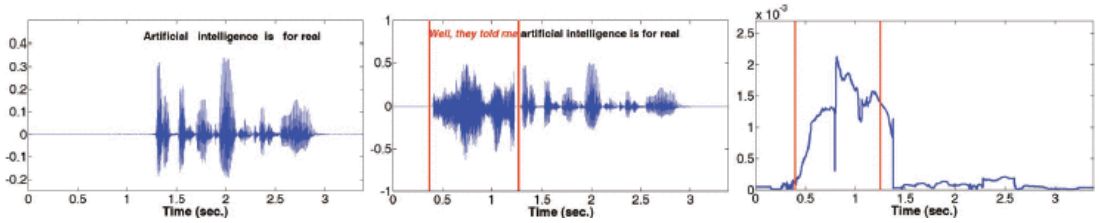
Kurtoza (*eng. "Kurtosis"*) je statistička mjera koja opisuje oblik distribucije. Konkretno, mjeri "skraćenost" ili "vrh" distribucije vrijednosti u zvučnom signalu. Distribucija s visokom

kurtozom ima oštriji vrh i deblje repove, što znači da ima ekstremnije vrijednosti. U kontekstu audio signala, visoka kurtoza mogla bi označavati zvuk s više prolaznih ili iznenadnih praska zvukova visoke amplitude. S druge strane, distribucija s niskom kurtozom je zaobljenija na vrhu i ima svjetlije repove, što znači da ima manje ekstremnih vrijednosti. U audio signalu, niska kurtoza može označavati zvuk koji je dosljedniji i ima manje iznenadnih praska zvuka. Kurtoza može biti korisna u audio analizi za identificiranje značajki ili karakteristika zvučnog signala koje možda nisu vidljive jednostavnim vizualnim pregledom valnog oblika. Na primjer, može pomoći u prepoznavanju buke, otkrivanju anomalija ili klasificiranju različitih vrsta audio signala. [48]

TIMIT korpus čitanja govora osmišljen je kako bi pružio govorne podatke za akustično-fonetske studije te za razvoj i evaluaciju sustava automatskog prepoznavanja govora. TIMIT sadrži širokopojasne snimke 630 govornika osam glavnih dijalekata američkog engleskog jezika, od kojih svaki čita deset fonetski bogatih rečenica. TIMIT korpus uključuje vremenski usklađene ortografske, fonetske i transkripcije riječi, kao i 16-bitnu datoteku govornog valnog oblika od 16 kHz za svaku izreku. [49]

Autori Pan, Zhang i Lyu su definirali metodu koja je sposobna pronaći krivotvorene dijelove u audio zapisu na temelju procjene lokalne razine buke. Upotrebom audio kurtoze, mjere vrha distribucije, za procjenu diskretne kosinus transformacije (*eng. "Discrete Cosine Transform"* DCT) odgovora audio uzorka iz TIMIT skupa podataka. Nakon konvolucije sa 63 DCT AC filtra, vrijednosti odgovora kurtoze izračunavaju se i sortiraju, otkrivajući da se većina vrijednosti grupira oko srednje vrijednosti, osim nekoliko netipičnih. Taj se uzorak zatim koristi kao osnova za razvoj učinkovite metode za procjenu varijance buke dodane čistim audio signalima. Idući korak je globalna procjena razine buke koja se temelji na bliskoj postojanosti vrijednosti kurtoze u domeni Diskretne kosinus transformacije. Cilj je procijeniti varijancu aditivne bijele Gaussove buke (AWGN) od kontaminiranog signala, koristeći odnos između kurtoze izvornih i kontaminiranih signala i njihovih varijanci. Unatoč pretpostavci AWGN-a, doći će do miješanja ne-Gaussove neovisne buke zbog teorema središnje granice i neovisnosti o buci. S prethodna dva koraka metoda je već funkcionalna za testiranje nekog scenarija sljedeći korak je samo proširenje za efektivnije pronalaženje krivotvorina. Posljednji korak govori o proširenju metode za globalnu procjenu razine buke tako što predlaže učinkovit algoritam koji koristi dinamičko programiranje i koncept integralnog vektora, koji je prikaz audio signala gdje vrijednost na određenom indeksu predstavlja zbroj svih vrijednosti uzorkovanja do tog indeksa. Ovaj integralni vektor se zatim koristi za učinkovito izračunavanje lokalne varijance i kurtoze u audio signalu, pomažući u identifikaciji lokalne razine buke i potencijalne krivotvorine zvuka. [50]

Definirano je nekoliko scenarija za testiranje ove metode, a jedan od njih je taj da su uzeli određeni audio zapis iz skupa podataka TIMIT koji glasi "umjetna inteligencija je stvarna" (*eng. "artificial intelligence is for real"*. Ovom zapisu je dodan metodom spajanja (*eng. "Splicing"*) glasovni zapis "Pa rekli su mi" (*eng. "Well, they told me"* na početak prethodnog glasovnog zapisa. Kako bi na kraju cijeli audio zapis zvučao "Pa rekli su mi umjetna inteligencija je stvarna". U nastavku je prikazana (slika 13) detekcije mjesta manipulacije u audio zapisu. [50]



Slika 13: Prikaz detekcije audio krivotvorine spajanja (*eng. "Splicing"*) na temelju procjene lokalne razine buke (Izvor: Pan, Zhang i Lyu, 2012)

3.2.7. Strojno učenje i umjetna inteligencija

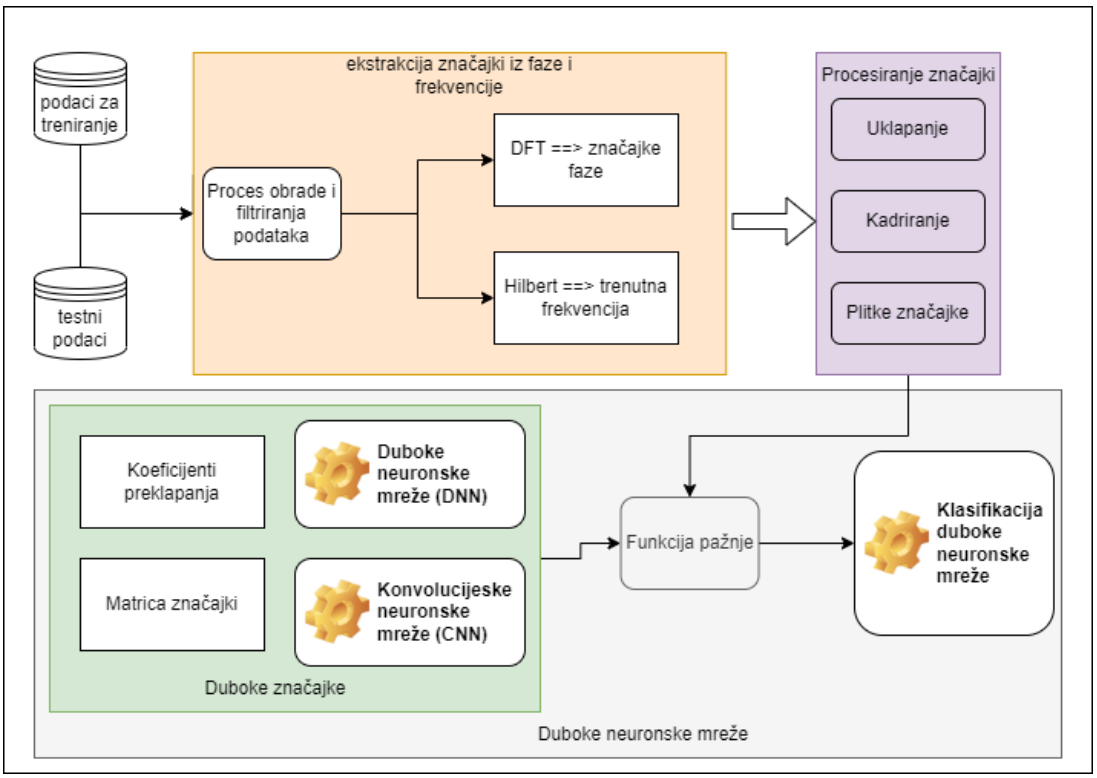
Razvoj algoritama strojnog učenja i umjetne inteligencije omogućava automatsko otkrivanje nepravilnosti i falsifikata u glasovnim zapisima. Tehnike dubokog učenja, poput neuronskih mreža, mogu se koristiti za prepoznavanje obrazaca koji su specifični za manipulaciju zvuka.

3.2.7.1. Fuzijom plitkih i dubokih značajki za pronalaženje krivotvorina

Wang, Yang, Zeng i dr. su definirali metodu pronalaženja krivotvorina na temelju neuronskih mreža i dubokog učenja tako što radimo fuziju plitkih i dubokih značajki. Metoda je podijeljena na sljedeće korake: ekstrakcija značajki iz faze i frekvencije, procesiranje značajki i duboka neuronska mreža. Ekstrakcija značajki faze i frekvencije filtrira i obrađuje audio zapise koji se kasnije profilira s diskretnom Fourier transformacijom (*eng. "discrete Fourier transform"* DFT) iz koje dobivamo vrijednosti značajke faze. Primjenom Hilbertove transformacije na filtrirani signal dobivamo trenutnu frekvenciju. Hilbertova transformacija se koristi zato što olakšava formiranje analitičkog signala koji potreban prilikom analitičke obrade glasa. Procesiranje značajki je idući korak u kojem se koriste prethodno dobivene vrijednosti za trenutnu frekvenciju i značajku faze. Koristimo srednje vrijednosti faze frekvencije elektronske mreže (*eng. "electronic network frequency "* ENF) i trenutne pomake frekvencije kao plitke značajke, a zatim se koristi konvolucijska neuronska mreža za učinkovitije učenje zamršenih detalja ovih faza i frekvencija ENF-a, čime dobivamo duboke značajke. Podatke o fazi i frekvenciji ENF-a obrađujemo uokvirivanjem, preoblikovanjem i uklapanjem kako bi se koristili kao ulaz za neuronsku mrežu, što zatim pruža duboke značajke potrebne za fazu obuke mreže. U posljednjoj fazi neuronska mreža dobiva matricu značajki i koeficijente preklapanja kao ulazne podatke kako bi kao izlazni produkt dobila značajke koje sadrže istovremeno lokalne i globalne informacije. Da bi izvršili spajanje plitkih i dubokih značajki koristimo mehanizam pažnje, a sam proces počinje ulančavanjem značajki faze i frekvencije kako bi se dobio ulaz duljine (Ł) . Zatim, da bi se odredila težina svake značajke, ulaz prolazi kroz potpuno povezani sloj koristeći dvije aktivacijske funkcije, ReLU za poboljšanje nelinearnosti i Sigmoid za dobivanje težine, koja se zatim množi s ulaznim značajkama. U ovoj studiji, mehanizam fuzije pozornosti koristi Sigmoidnu aktivacijsku funkciju za dobivanje težina, budući da je glavni cilj potisnuti nevažeće značajke, a ne optimiziranje istih. Takav mehanizam može automatski dodijeliti različite težine svakoj značajki iz plitkih i dubokih slojeva, dajući veće težine značajkama koje značajno utječu na rezultat klasifikacije i manje težine manje utjecajnim značajkama, čime se povećava točnost otkrivanja krivotvorina i

generalizacija modela.[40]. Cijeli proces je grafički prikazan na (slici 14).

Za bolje razumijevanje definirat ćemo funkciju pažnje koja mapira upite i skupove od para ključ-vrijednost i izlaza , gdje su upiti, ključevi, vrijednosti i izlaz vektori. Vrijednost izlaza kalkuliramo ponderiranim zbrojem vrijednosti, gdje se težina dodijeljena svakoj vrijednosti izračunava pomoću funkcije kompatibilnosti upita s odgovarajućim ključem.[51]



Slika 14: Prikaz procesa fuzije plitkih i dubokih značajki pomoću neuronskih mreža i dubokog učenja (Izvor: Wang, Yang, Zeng i dr., 2022)

3.3. Detekcija krivotvorina sinteze govora (*eng. "Speech synthesis"*)

Postoji nekolicina metoda koje se bave detekcijom sinteze govora no ja sam izdvojio dva rada koja su usko povezana ali se ipak razlikuju jer se drugom metodom ekstrahiranja značajki dolazi do zaključka o pitanju krivotvorenih glasova. AlBadawy, Lyu i Farid su kreirali metodu koja pomoću bispektralne analize uspijeva detektirati govor koji je konstruiran s umjetnom inteligencijom. Za testiranje ove metode korišteni su glasovni zapisi kreirani od 6 web aplikacija koje pretvaraju tekst u govor sintezom govora (Amazon Polly, Google WaveNet, GAN, Lyrebird, Baidu). Druga metoda je definirana od strane Singh i Singh koji uz bispektranlnu analizu rade i Mel kepstralnu analizu (*eng. "Mel Frequency Cepstral Coefficient"* MFCC) za dobivanje dodatnih komponenata iz ljudskog glasa, kakve umjetna inteligencije ne može proizvesti.

Ekstrakcija značajki započinje tako da se audio signal prvo rastavlja pomoću Fourierove transformacije, zatim se spektar snage koristi za otkrivanje korelacija drugog reda. Međutim, spektar snage ne može otkriti korelacije višeg reda, koje su od najvećeg interesa. Ovo je

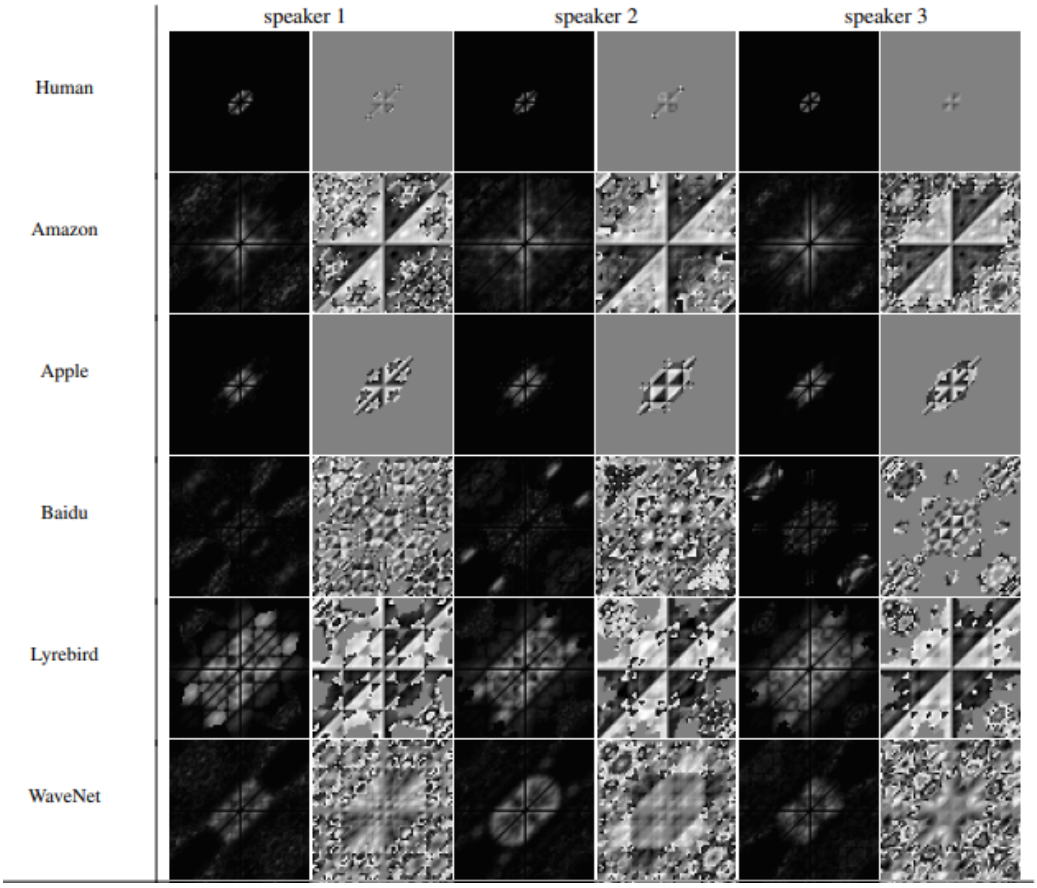
mjesto gdje bispektar stupa na scenu, omogućujući otkrivanje korelacija trećeg reda i otkrivajući korelacije između različitih harmonijskih tripleta. Bispektar je veličina kompleksnih vrijednosti, izražena kao magnituda i faza, te normalizacijom se dobivaju vrijednosti između [0, 1]. U prisustvu šuma, kako bi se osigurale stabilne procjene, signal se dijeli na više segmenata, a bikoherencija (normalizacija bispektra) se izračunava na temelju prosjeka spektra bikoherencije svakog segmenta. U nastavku su prikazane formule za računanje magnitude bikoherencije (normalizacija bispektra) (formula 2) i faza bikoherencije (formula 3.[52], [53])

$$\left|\hat{B}\left(\omega_1,\omega_2\right)\right| = \frac{1}{K}\sum_{K}\left(\left|Y_K\left(\omega_1\right)\left\|Y_K\left(\omega_2\right)\right\|Y_K\left(\omega_1+\omega_2\right)\right|\right) \tag{2}$$

$$\angle\hat{B}\left(\omega_1,\omega_2\right) = \frac{1}{K}\sum_{K}\left(\angle Y_K\left(\omega_1\right)+\angle Y_K\left(\omega_2\right)-\angle Y_K\left(\omega_1+\omega_2\right)\right) \tag{3}$$

Za dobivanje Mel Kepstralnog koeficijenta (MFCC) autori Singh i Singh izračunavaju iz spektra magnitude kratkotrajne Fourierove transformacije audio signala, koji je podijeljen u segmente koji se preklapaju, a svaki je podvrgnut Fourierovoj transformaciji da bi se dobio spektar snage. Nakon primjene Mel frekvencijskog filtra na spektar snage, uzima se diskretna kosinus transformacija MFCC logaritma snage, a MFCC koeficijenti predstavljaju amplitudu rezultirajućeg spektra. Dodatne značajke korisne za razlikovanje govora uključuju Δ-Cepstrum i Δ2-Cepstrum, koje predstavljaju promjene u MFCC koeficijentima odnosno Δ-Cepstrum vrijednostima, a zajedno s MFCC-ovima pružaju robusne karakteristike za Kepstralnu analizu.[53]

Obje metode definiraju klasifikaciju modela za strojno učenje kako bi algoritam mogao razvrstati ljudski glas od glasova koji su kreirani sintezom govora. AlBadawy, Lyu i Farid su svoju podijelili 7 kategorija (više od 1800 glasovnih zapisa kreiranih sintezom govora i 100 ljudskih) [52], a autori Singh i Singh su koristili nešto manje audio zapisa (400 audio zapisa kreiranih sintezom govora i 250 ljudskih audio zapisa) koji su kategorizirani na 4 vrste [53]. Obje metode postižu visoku preciznosti pronalaska sintetičkih audio zapisa unatoč dodatnom manipulacijom audio zapisa (dodavanje buke, ponovnom kompresijom). Kao dodatan primjer u nastavku je prikazan (slika 14) grafički prikaz na kojem možemo vidjeti razliku između ljudskog glasa i glasa sinteze nakon izračuna normalizacije bispektra magnitude i faze. Možemo primijetiti jedinstveni otisak ljudskog glasa usporedno s pet sintetičkih glasova koji su kreirani od istih glasova pomoću sinteze govora.



Slika 15: Grafički prikaz normalizacije bispektra magnitude i faze za ljudski govor i 5 različitih govora sinteze.Stupci prikazuju 3 različita ljudska glasa, redovi prikazuju izvor glasa, gdje je prvi originalni ljudski glas, a ostali su kreirani pomoću sinteze govora od istih glasova. (Izvor: AlBadawy, Lyu i Farid, 2019)

4. Praktični rad

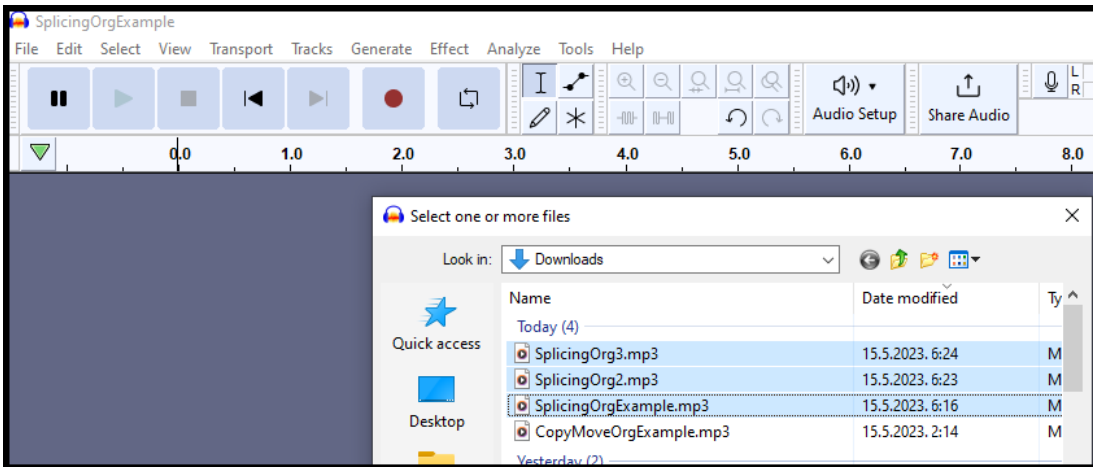
Praktični rad će se sastojati od jednog scenarija gdje ćemo pomoću spektrograma pokušati analizirati krivotvorene audio zapise s originalnim. Analiza se će se sastojati od računanja MFCC koeficijenata i prikaz na spektrogramu, izračun amplitude glasa i prikaz, uz svaki prikaz će se paralelno prikazivati originalni zapis i krivotvoreni. Sljedeće krivotvorine će biti korištene: kopiraj, premjesti(*eng. "Copy-move"*), spajanje (*eng. "Splicing"*) i sinteza govora (*eng. "speech synthesis")*. Za kreiranje krivotvorenog zapisa metodom kopiraj-premjesti i spajanje korišten je softver Audacity, cijeli proces kreacije će biti detaljno opisan i popraćen slikama. Sinteza govora je kreirana s web aplikacijom Resemble.AI, postupak za kreiranje umjetnog glasa će biti u nastavku dokumentiran. Preostaje nam dio za analizu koji je izveden u programskom jeziku Pythonu, koja će na temelju koda i prikaza spektrograma biti protumačena.

4.1. Kreiranje krivotvorina

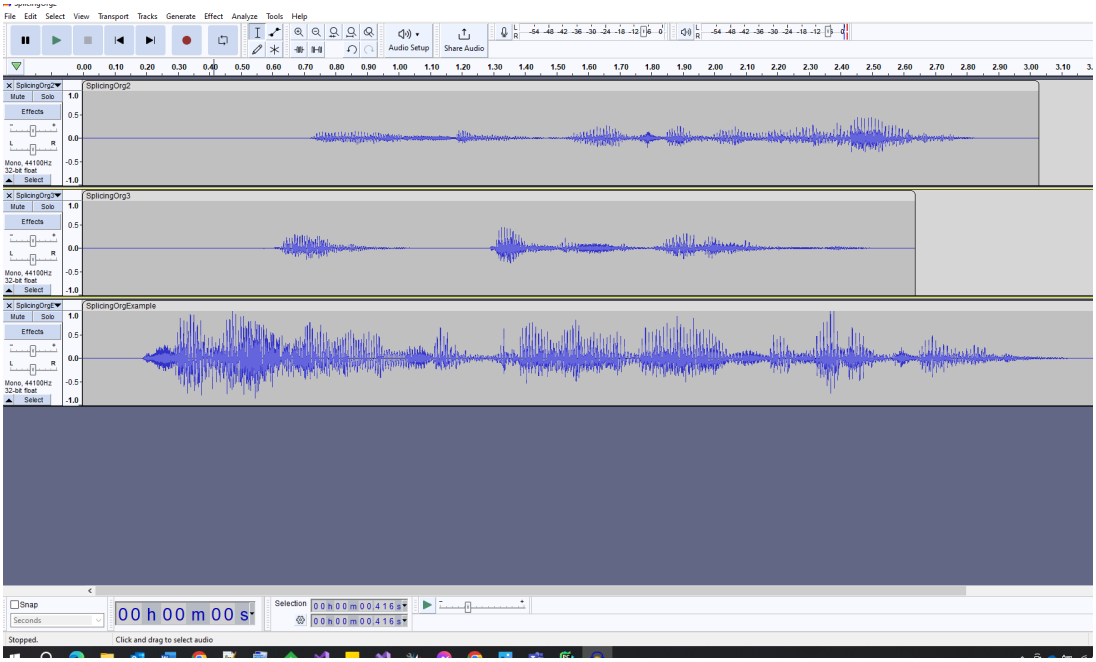
Audacity je korišteni softver za kreiranje krivotvorenih glasovnih zapisa metodama kopiraj-premjesti i spajanje korišten , vrlo detaljno ćemo odraditi postupak kreiranja sami krivotvorina te ujedno i malo bolje približiti korištene alate i njihove mogućnosti. Sinteza govora je kreirana s web aplikacijom Resemble.AI, postupak za kreiranje umjetnog glasa s dubokim učenjem i neuronskim mrežama će biti ukratko objašnjen.

4.1.1. Spajanje (*eng. "Splicing"*)

Kako bi mogli kreirati krivotvorinu potrebno je uvesti željene audio zapise u aplikaciju (slika 16). Kao kratak uvod malo ćemo opisati sučelje aplikacije za lakše praćenje u nastavku rada (slika 17). Sučelje je vrlo jednostavno dizajnirano i lako za koristiti, pri vrhu se nalazi glavni izbornik iz kojeg možemo glavne funkcionalnosti testirati kao što su pokretanje, pauziranje, klikom na spektrogram odabiremo početka pokretanja audio zapisa, mijenjanje glasnoće glasovnog zapisa i mnogo drugih. Odmah možemo vidjeti kako su naši glasovni zapisi lijepo strukturirani jedan ispod drugog, a ujedno alat prikazuje valni oblik audio zapisa pomoću grafa za lakšu manipulaciju istog. Svaki audio zapis s lijeve strane sadrži zaseban izbornik na kojem se izmjene direktno primjenjuju na audio zapis. Mijenjanje glasnoće zapisa, dodavanje efekata (reverberacija, distorzija i drugi), pomicanje zvuka u željenu stranu.

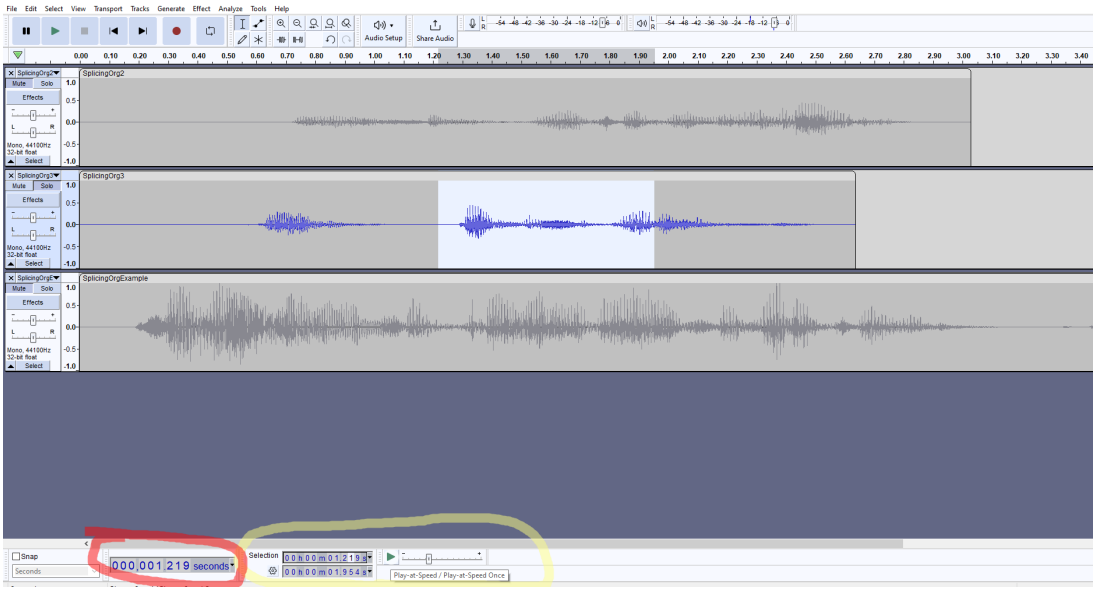


Slika 16: Uvoz audio zapisa u Audacity aplikaciju

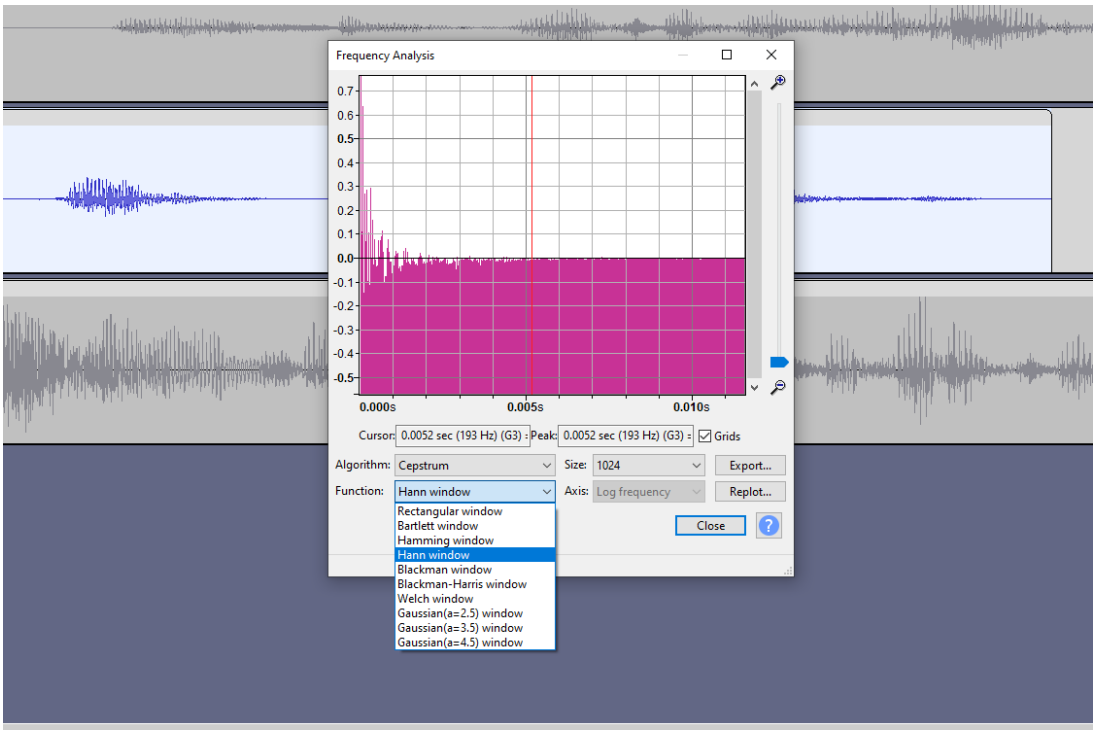


Slika 17: Prikaz uvezenih audio zapisa i pregled samog sučelja aplikacije

Prije nego što krenem s opisivanjem postupka kreiranja krivotvorina, razmotrio bih (slika 18) podnožje aplikacije. Žutom bojom označene su dvije funkcije od kojih jedna mijenja tempo audio zapisa, a druga služi za unos vremena od-do kojeg želimo označiti, pa zatim uređivati samo označeni odjeljak kao što je prikazano na slici. Na srednje spektrogramu možemo vidjeti kakao izgleda označeni dio, crvenim markerom je označeno vrijeme koje prikazuje na kojoj minuti je odabran audio zapis. Vrijedno spomena je još da ovaj alat ima mogućnost kreiranja spektrograma na temelju funkcija Hamming window, Gaussian i drugih. Ostataka funkcija možemo vidjeti na (slici 19).



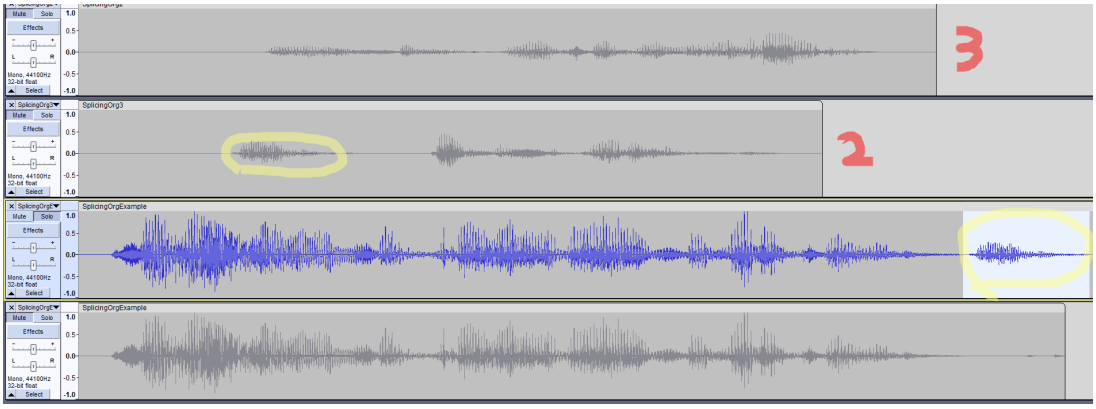
Slika 18: Prikaz podnožja aplikacije s istaknutim elementima vremena i tempa



Slika 19: Prikaz spektrograma analize i mogućnost odabira željene funkcije

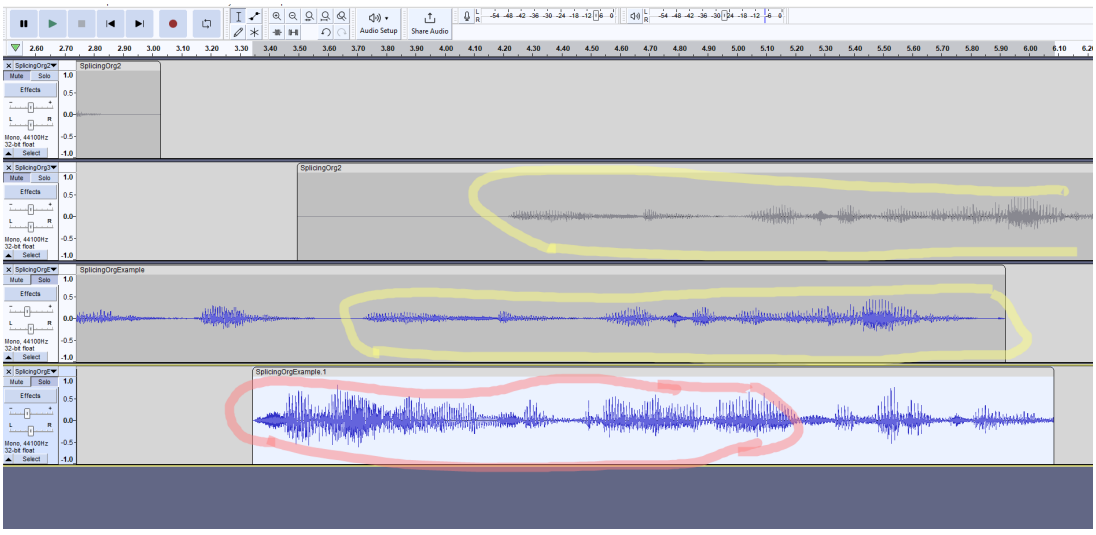
Snimljene su 3 glasovne poruke s mobilnom aplikacijom za kreiranje glasovnih poruka. Kako bih konstruirao spajanje napravljene su 3 snimke sa sljedećim kontekstima. Prva snimka će biti baza na koju ćemo dodavati ostale odnosno dio teksta koji želimo. Kontekst prve snimke je sljedeći "Cijelu noć sam proveo za računalom". Druga snimka će biti označena s brojem 2 i kontekst joj glasi "I tako se to radi!" iz koje ćemo uzeti jednu riječ a to je "i", koju ćemo dodati na kraj prve snimke. Treća snimka će biti označena brojem 3 i njen kontekst je sljedeći "Nisam jučer imao vremena", cijeli zapis ću spojiti s prvom snimkom tako da bi krajnji produkt trebao ovako zvučati " Cijelu noć sam proveo za računalom i nisam jučer imao vremena". (slika 20)

prikazuje postupak spajanja dijela druge snimke u prvu, kako bi dobili ovaj rezultat označimo dio audio zapisa kao što je žutim markerom na slici prikazano na audio zapisu broj 2, kopiramo odlomak s desnim klikom i odabirom kopiranje, te prebacimo dio na željeni audio i mjesto na kojem hoćemo. Možemo sada već primijetiti kako audio zapisi broj 2 i 3 imaju manje amplitude što govori da su dosta tiše od one u koju će biti spojene. Što će predstavljati problem kada se bude izvodila audio snimka jer će prvi dio biti glasan a drugi jako tih.

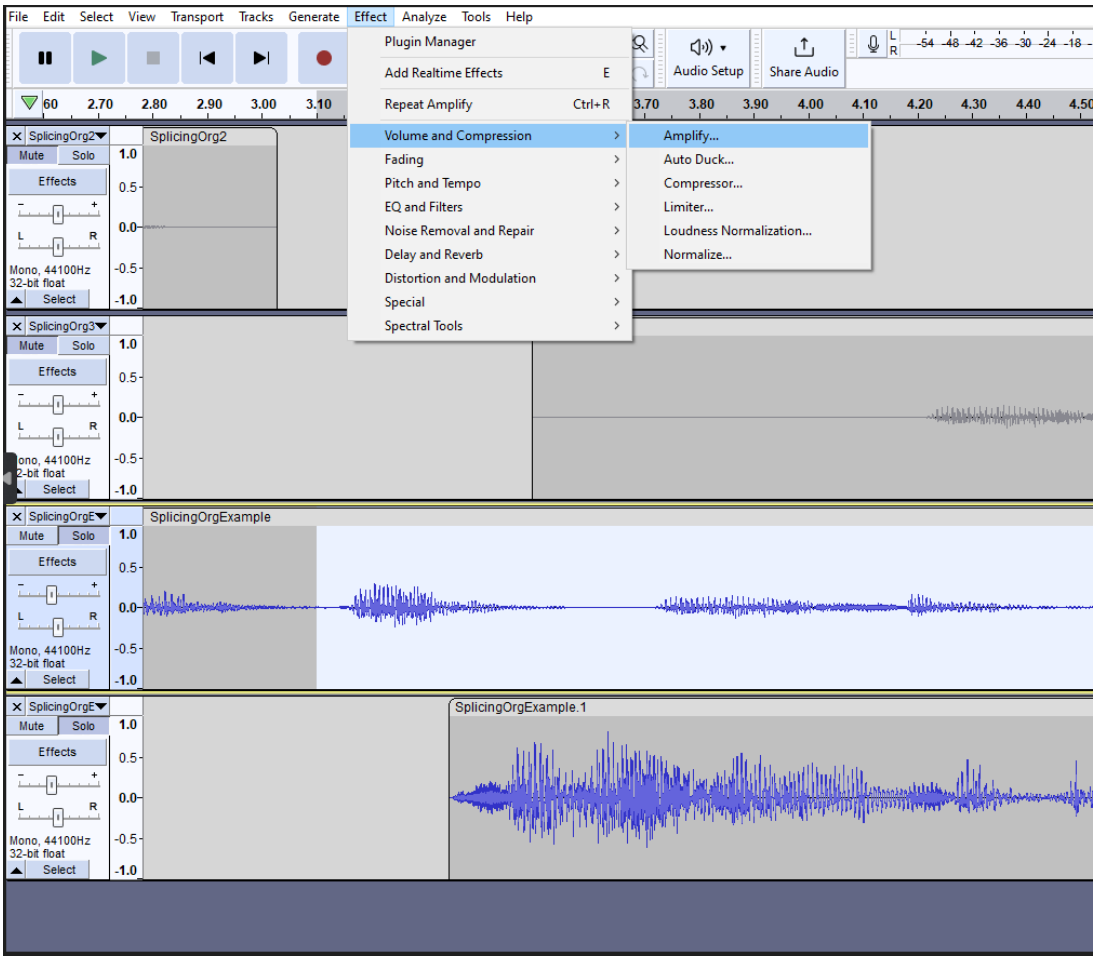


Slika 20: Prikaz spajanja snimke broj 2 u snimku 1 (dodavanje riječi "i" na kraju prve snimke)

Kako bi dovršili spajanje označavamo snimku broj 3 i dodajemo ju na kraj snimke broj 1 što je prikazano na (slici 21) gdje faza spajanja prikazan žutim markerom. Trenutno naša snimka zvuči ovako "Cijelu noć sam proveo za računalom i nisam jučer imao vremena" Crveni marker označava problem neujednačene glasnoće između audio zapisa, u kojem se vidi jasno a i čuje da je spojeni dio u snimci puno tiši. Kako bi riješili ovaj problem moramo smanjiti amplitudu na prvom dijelu snimke i povećati na drugom dijelu (slici 22).

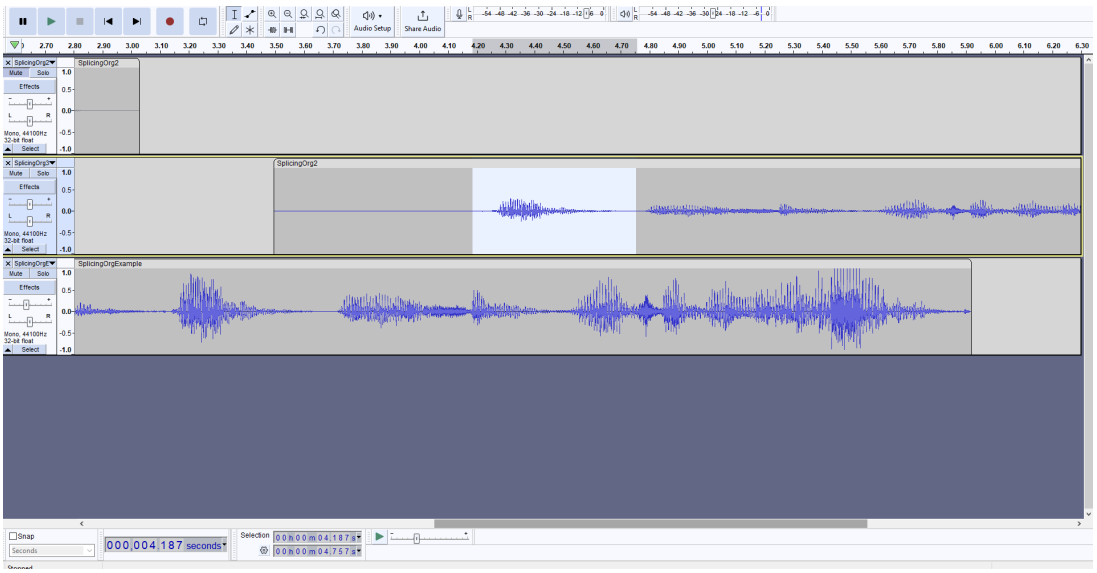


Slika 21: Prikaz spajanja snimke 3 u snimku 1 i prikaz problema nejednake glasnoće audio zapisa



Slika 22: Proces povećanja amplitude za novo dodane dijelove glasovne poruke

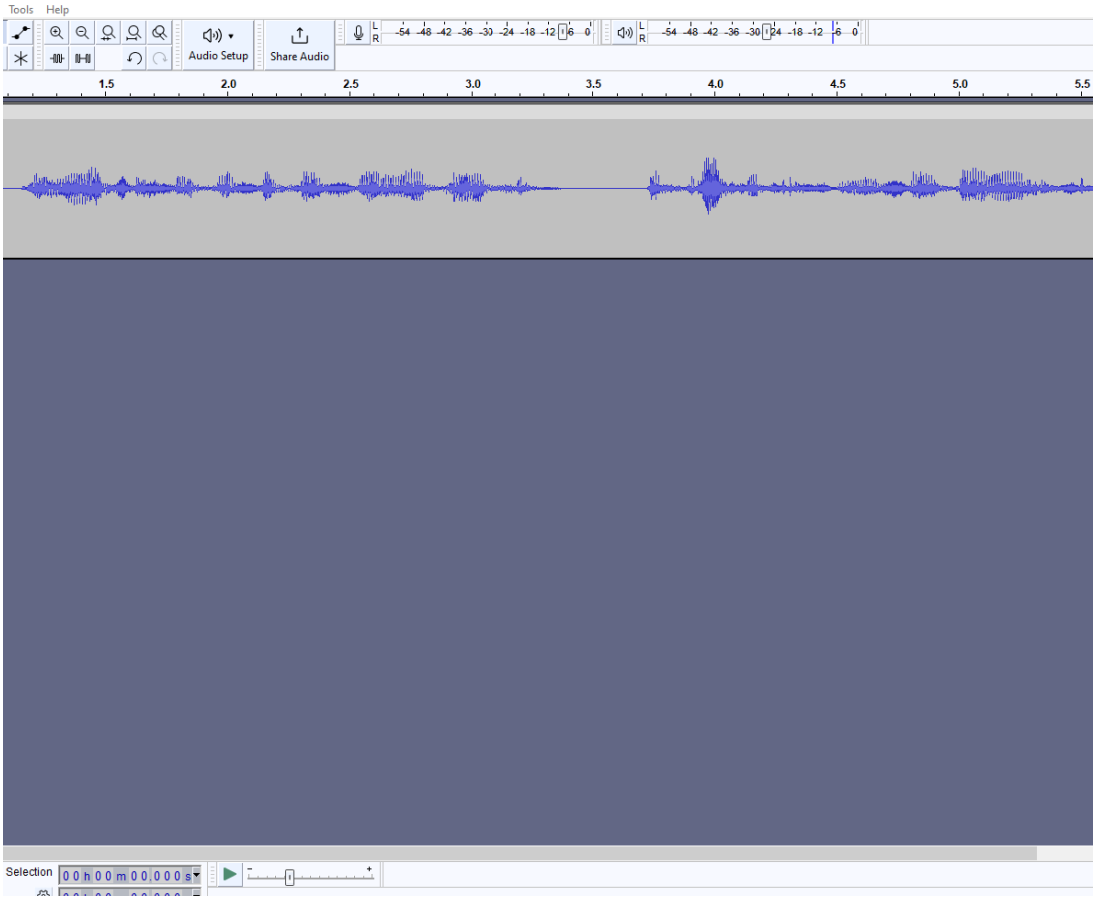
Nakon što smo obavili proces amplifikacije, dodavanjem 9 dB novo pripojenom dijelu audio zapisa, znatno se vidi povećanje amplituda na valnom obliku što je i prikazano na (slici 23).



Slika 23: Usporedba audio zapisa prije amplifikacije i poslije

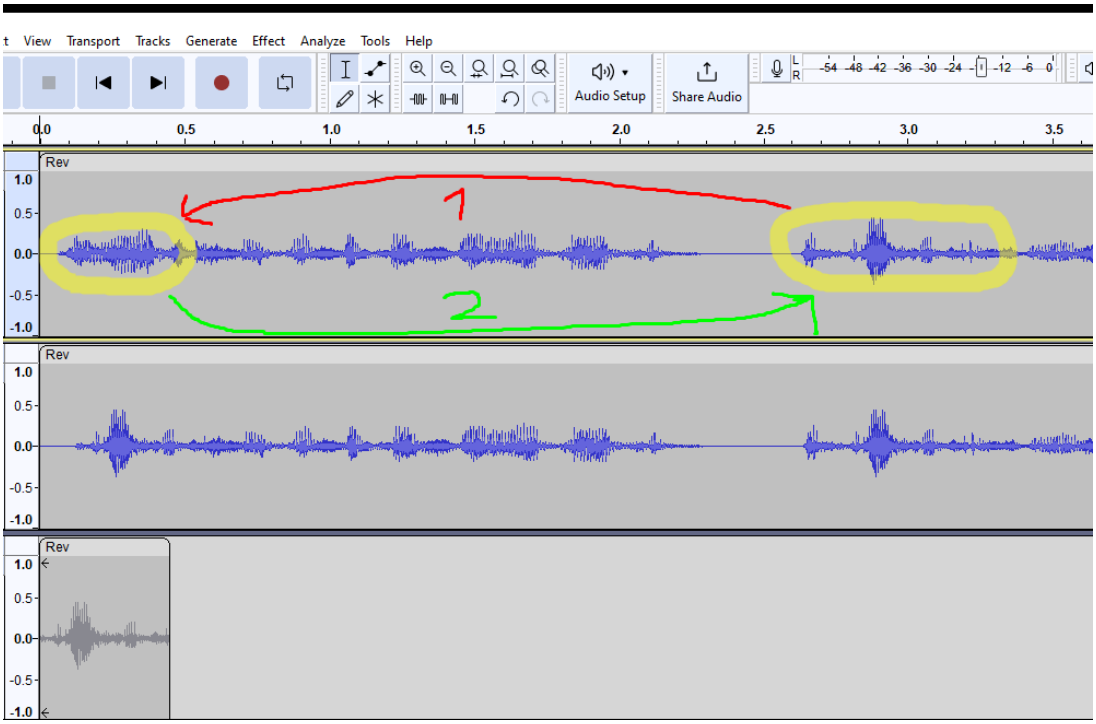
4.1.2. Kopiraj, premjesti(*eng. "Copy-move"*)

Kako smo u poglavlju prije već definirali postupak spajanja, kopiraj, premjesti će biti puno jednostavniji jer je princip dosta sličan ali malo zahtjevniji jer je potrebna preciznost. Kreiran je jedan audio zapis jer kopiraj, premjesti metoda koristi samo dijelove vlastitog glasovnog zapisa i premješta ih unutar sebe, tako da se kontekst izmijeni. Snimljena je glasovna poruka kao i u prethodnom scenariju za spajanje, kontekst poruke je sljedeći "Jučer sam posjetio svoju obitelj, a prekjučer sam išao tom ulicom", izgled spektrograma prikazan je na (slici 24).



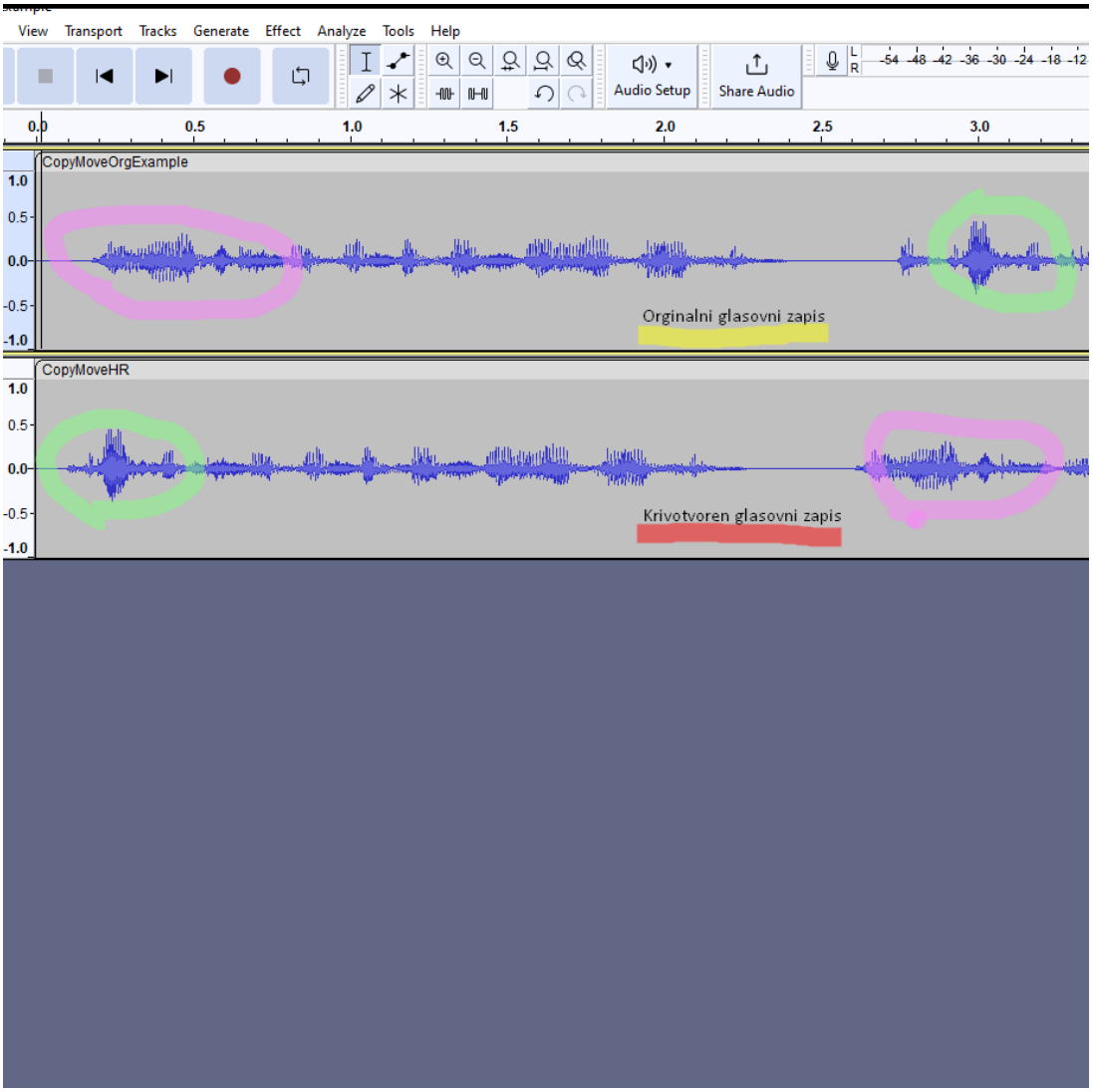
Slika 24: Spektrogram prikaz glasovnog zapisa

Ideja za kreiranje krivotvorine je zamijeniti riječima "Jučer" i "prekjučer" mjesta tako da bi poruka trebala glasiti "Prekjučer sam posjetio svoju obitelj, a jučer sam išao tom ulicom". Ako pretpostavimo da je glasovna poruka uzeta kao dokaz za suđenje nekoj osobi onda doista igra ulogu kada si i gdje si bio. U nastavku je prikazana (slika 25) na kojoj možemo vidjeti proces zamjene gore navedenih riječi kako bi se kontekst manipulirao. Na slici vidimo da su žutom bojom označene riječi "Jučer" i "prekjučer" koje želimo premjestiti. U drugom retku već vidimo da je riječ "prekjučer" premještena na početak rečenice, što znači da još trebamo prebaciti i doraditi drugi dio.



Slika 25: Detaljni Proces realizacije kopiraj premjesti metode

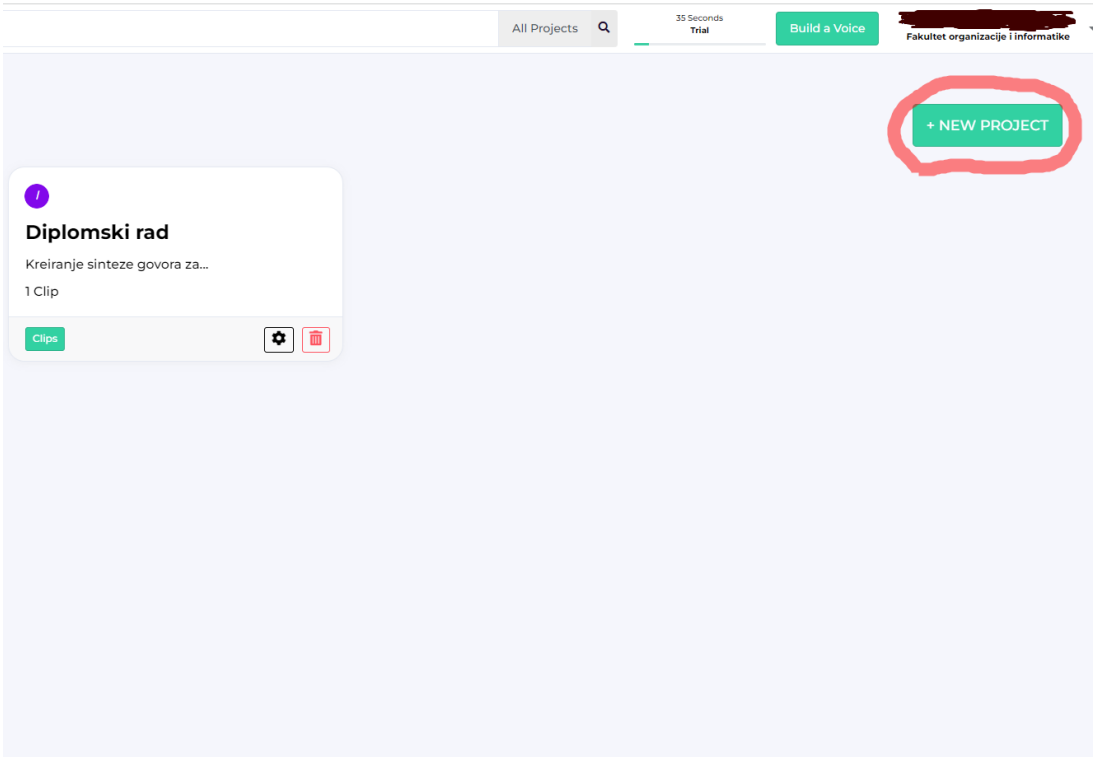
Premještanje riječi "Jučer" je bio malo zahtjevniji posao gdje se moralo u milisekundu pogoditi umetanje riječi kako bi se uklopila između "a" i "sam išao tom ulicom". Razlog tome je što prilikom snimanja nisam pripazio na vremenski razmak između izgovorenih riječi, pa je zbog toga stanka prije i poslije izgovorene riječi "prekjučer" skoro pa nevidljiva. Potrebna je velika pažnja i preciznost opažanja kako bi se čulo završavanje fonema ili slova, da se ne bi prekid čuo nakon premještaja. Na (slika 26) vidimo novonastalu krivotvorinu pomoću metode kopiraj, premjesti. Zelena boja reprezentira riječ "Jučer" a rozna riječ "prekjučer" kako bi se odmah uočila krivotvorina.



Slika 26: Prikaz različitosti krivotvorine i originalne glasovne poruke

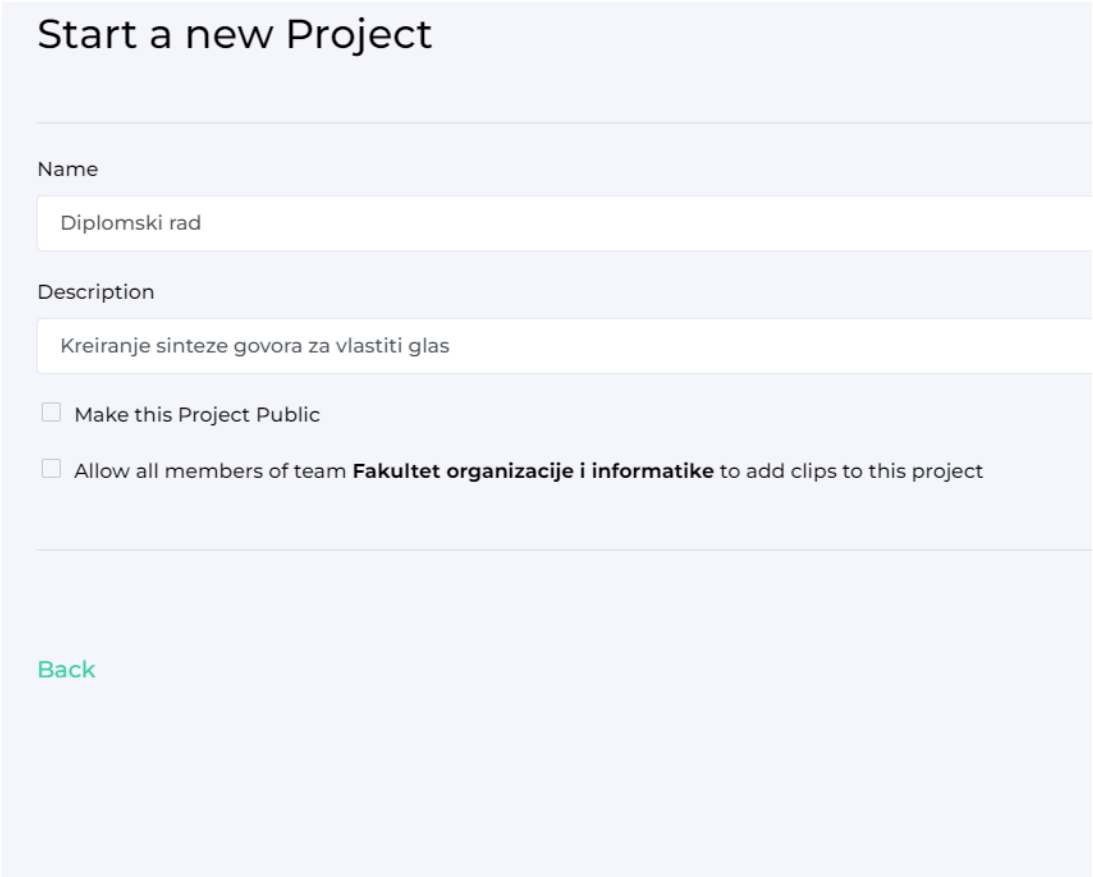
4.1.3. Sinteza govora ili tekst u govor (*eng. "Speech synthesis"*)

Krivotvorina sinteze govora je kreirana uz pomoć Resemble.ai aplikacije koja generira novi glas na temelju neuronskih mreža i dubokog učenja. Zbog porasta audio deepfake krivotvorina zadnjih par godina, htio sam vidjeti koje se sličnosti i razlike mogu pronaći kada se usporedi s originalnim glasom. Nakon što sam pomoću aplikacije ressemble ai reproducirao vlastiti deepfake glas, bio sam ugodno iznenađen kako je dosljedno napravljen. Očaran rezultatom odmah sam poslao glasovnu poruku svom kolegi koji nije ni primijetio da se radi o deepfake glasu. Sami proces generiranja glasa traje oko 20-30 minuta jer aplikacija u tih nekoliko generira nasumične tekstove koje moraš pročitati u mikrofon kako bi dobio vjerodostojan umjetni glas. Na samom početku je opaska napisana, "Osoba treba biti blizu mikrofona" kako bi se uhvatio cijeli spektar glasa za bolji krajnji ishod. Sučelje Resemble.AI web aplikacije je vrlo jednostavno i ugodno za koristiti, možemo vidjeti na (slici 27). Jedina mana je ta što će govor biti na engleskom, jer su s besplatnom verzijom dostupni samo engleski, španjolski i francuski, dok se za korištenje ostalih mora plaćati mjesečna pretplata.



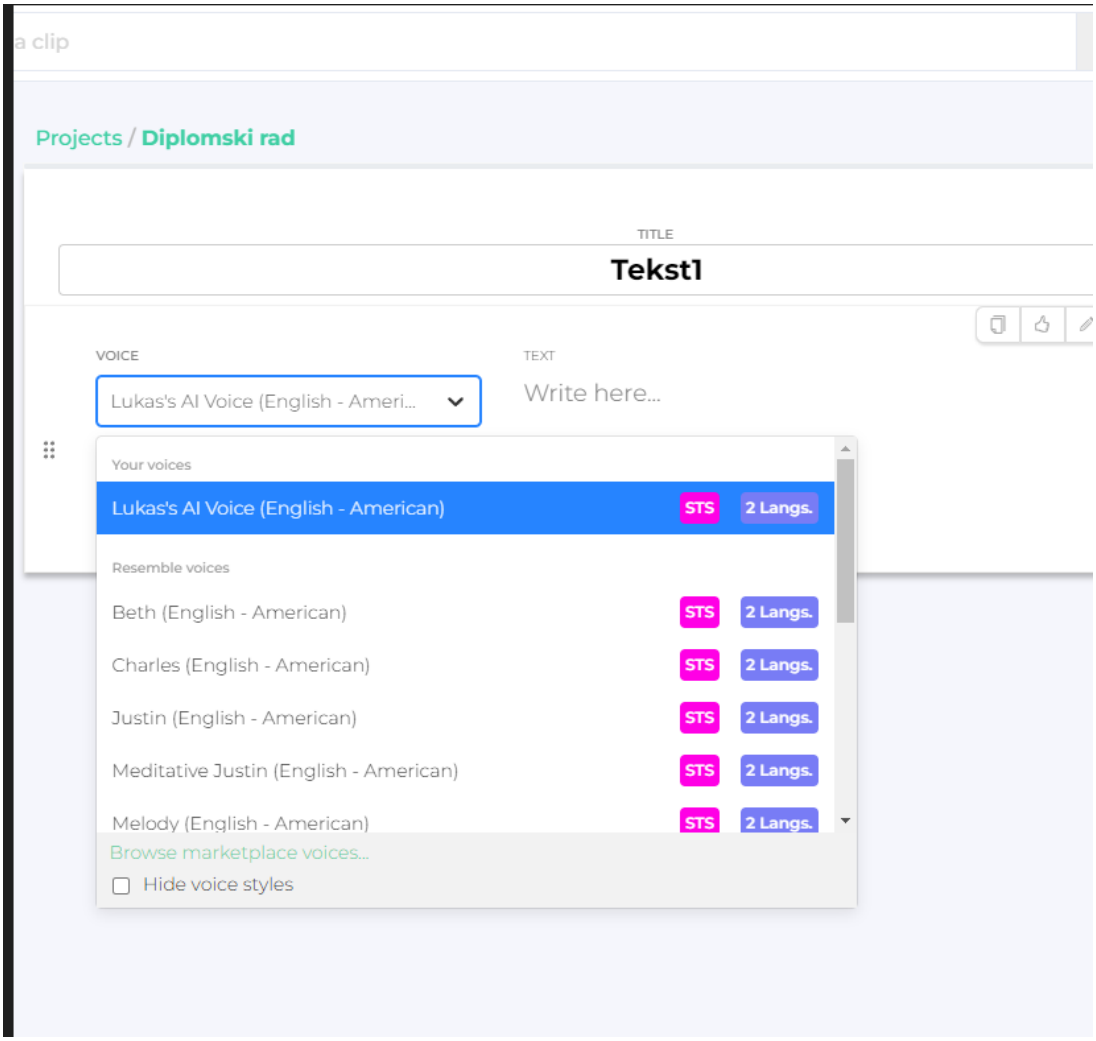
Slika 27: Prikaz sučelja Resemble.ai web aplikacije

Kako i sami vidimo da nema mnogo opcija na prvu tj. doista vrlo jednostavan dizajn s nekolicinom osnovnih funkcija. Za početak bi kliknuli na zeleni gumb koji se nalazi u desnom gornjem uglu, označeno je crvenim flomasterom. Kako bi kreirali vlastiti projekt u kojem možemo spremati i obrađivati nove audio zapise (slika 28). Potrebno je unijeti naziv i kratak opis za vaš projekt koji može ostati prazan jer nije obavezno polje. Dodatne opcije su da vam projekt postane javno dostupan za svakog, odnosno slobodna upotreba vašeg umjetnog glasa. Druga opcija nudi mogućnost pristupa svim članovima vaše organizacije u slučaju da označite polje i unesete validnu organizaciju. Nakon odabranog kreiramo projekt i tada možemo početi generirati glasovne poruke s vlastitim sadržajem.



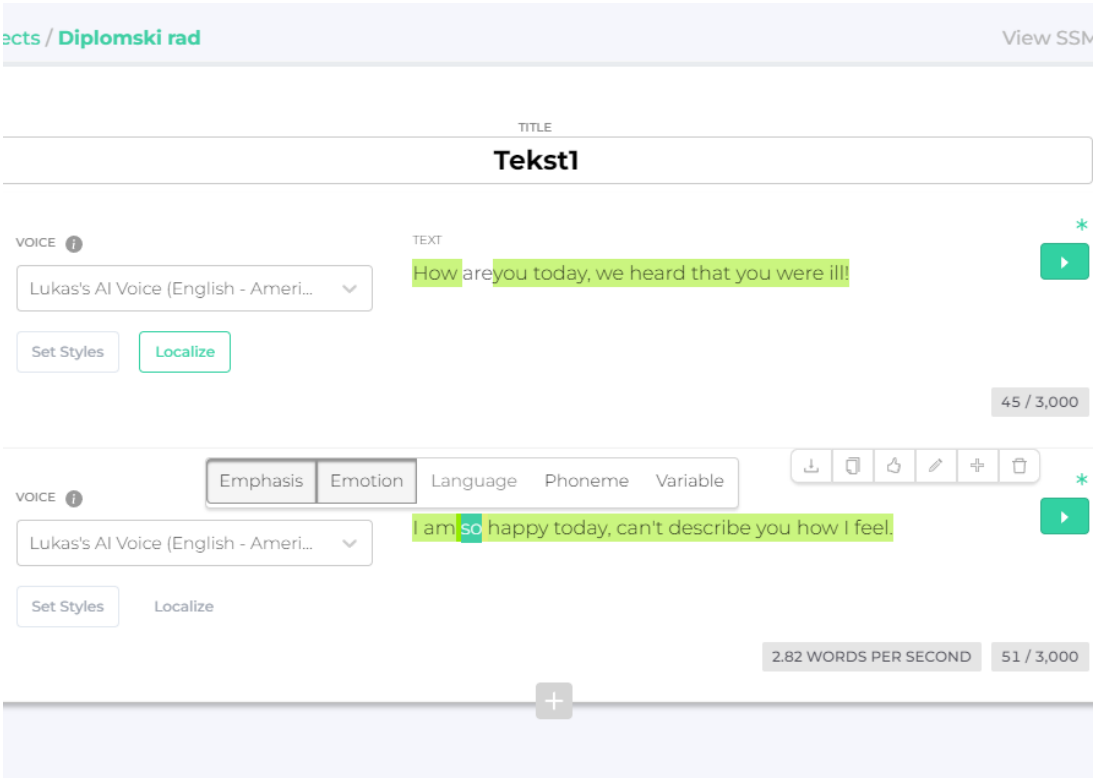
Slika 28: Postupak kreiranja projekta

Ulazimo u datoteku koju smo upravo kreirali tako da zanimljiviji dio sad kreće, definiranja teksta, efekata, jezika ako ste premium pretplatnik. Ako koristite besplatnu verziju tada ste uskraćeni za pojedine funkcionalnosti kao što su raznoliki jezici i kreiranje boljeg sintetičkog glasa uz veći i raznolikiji trening. Na (slici 29) vidimo s lijeve strane izbornik za odabir jezika i druge umjetne glasove koje možemo odabrati ujedno sa svojim. Odabirom željenog umjetnog glasa u kojeg će biti pretvoren tekst, možemo um prepustiti mašti i upisati proizvoljni tekst.



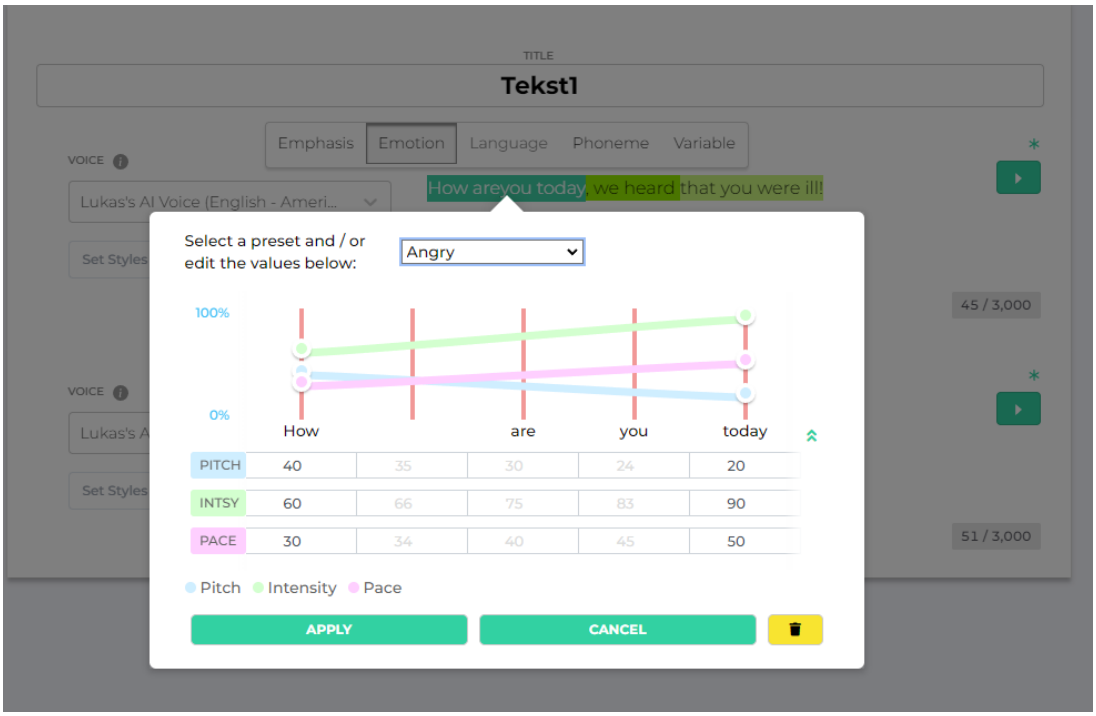
Slika 29: Postupak odabira sintetičkog jezika i unos teksta

Unosom teksta u polje i pokretanjem gumba se pokreće model neuronske mreže koji ubrzo generira tekst u govor s odabranim glasom. Uočavamo da se tekst može dodatno konfigurirati tako što mu se doda emocija ili naglasak na neku riječ. (slika 30) prezentira dva teksta koji su generirani s mojim osobnim glasom. Kako sam već spomenuo engleski jezik se mora koristiti za pretvorbu texta u govor. Prijevod prvog teksta je "Kako si danas, čuli smo da ste bili bolesni!" (*eng. "How are you today, we heard that you were ill!"*), prijevod drugog teksta je "Danas sam jako sretan, ne moga vam opisati kako se osjećam" (*eng. "I am so happy today, can't describe you how i feel."*)



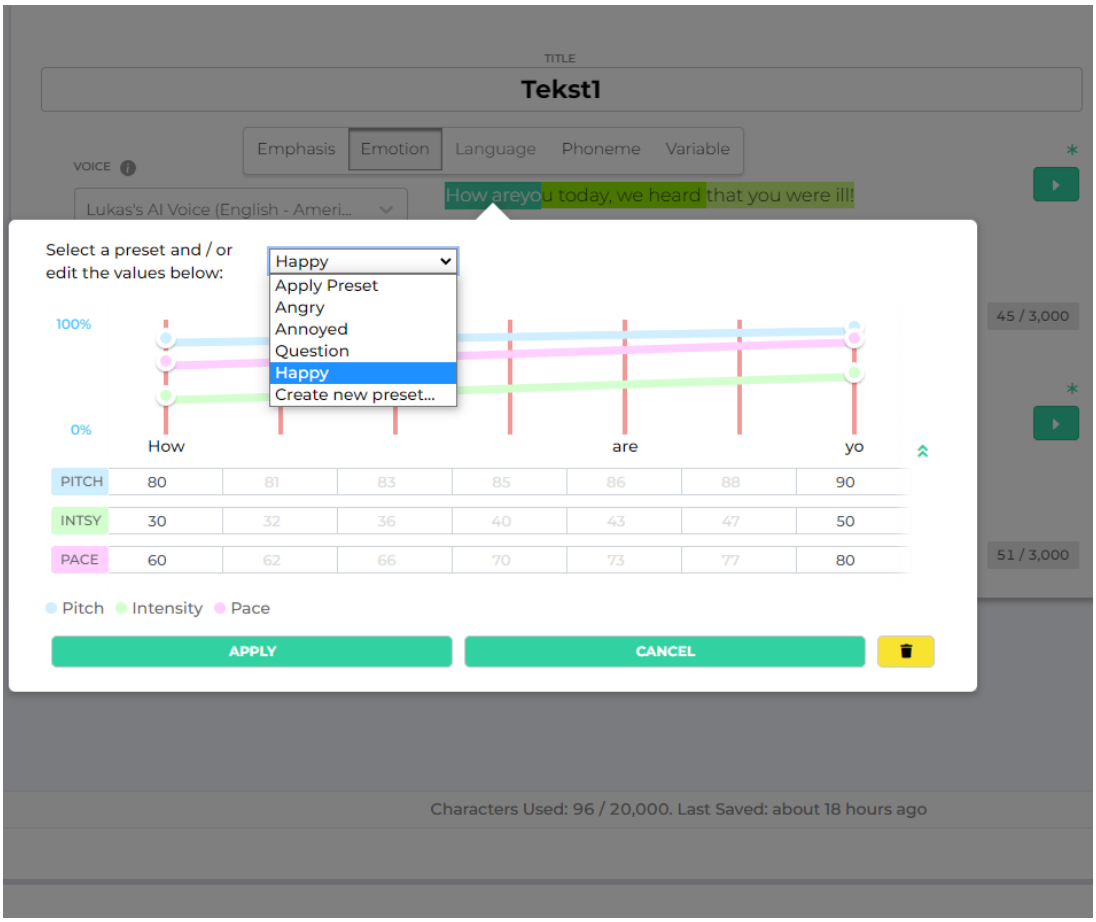
Slika 30: Generirani tekstovi i modificirani s efektima

Postoji mogućnost da se dodatno govor ukrasi, odnosno situira ako se želi prikazati emocija u govoru. Resemble.AI sadrži funkcionalnost koja može mijenjati emocije u govoru. To možemo ostvariti ako označimo tekst i odaberemo gumb za emocije otvara nam se novi izbornik na kojem možemo sami modificirati tempo, visinu tona i intenzitet (slika 31).



Slika 31: Odabir tempa, tona i intenziteta za svaku riječ

Odabrali smo u ovom primjeru emociju sreće gore u izborniku (slika 32), vidimo kako se skala na tonu i brzini povisila, slična reakcija kao što je u stvarnosti. Zbog osjećaja sreće osjećamo da glas postaje viši i govor je brži ako usporedimo s osjećajem ljutnje koji suprotnost. Testirao sam sve varijante ali nisam dobio dojam da je vrhunski izvedeno, pretpostavljam da se treba premium licenca uzeti kako bi se osjetila razlika jer ih model ne primjenjuje najbolje.



Slika 32: Odabir gotovih emocija

4.2. Python implementacije

Tijekom raspisivanja teorije dugo sam razmišljao što bih mogao kao praktični rad prezentirati za ovu impozantnu tematiku, što sam više istraživao dolazio sam do saznanja da je ova tema jako kompleksna. Kako je raslo saznanje kompleksnosti za ovu temu uvidio sam da nema brze i efikasne metode za pronalaženja krivotvorina u audio zapisima. Kako je ova grana još u usponu, i dosta metoda se još treba etablirati kako bi se mogao kreirati generalni algoritam koji će biti dostupan za široku publiku, ali mislim da ovakve stvari još dugo neće biti dostupne za široku publiku. Kako bih se vratio na vlastiti zaključak, što želim pokazati kao praktični primjer, a to je aplikacija koja će biti temeljena na pronalaženju krivotvorina pomoću vodenog žiga. A druga Python aplikacija će se temeljiti na metodi spektrogram analize, koja se u većini slučajeva koristi kao jedna uz neke druge metode. Jedan veliki razlog zašto sam uzeo ovu metodu je želja za dokazivanjem da i ljudsko oko može uvidjeti digitalnu manipulaciju kao

što je deepfake ili sintezu govora.

4.2.1. Digitalni žig implementacija

Metoda digitalnog žiga je već objašnjena ali ponovit ću princip rada ukratko. To je tehnika koja umetanjem skrivenih podataka na takav način da ju slušatelj ne može primijetiti, a istovremeno je robusna protiv izmjene samog signala i ne utječe na degradaciju kvalitete izvornog zvuka. Trebao bi preživjeti različite operacije obrade signala (poput kompresije, filtriranja ili pretvorbe), a istodobno neprimjetan kako ne bi degradirao kvalitetu izvornog zvuka. Trebalo bi se osigurati da metodu bude sigurna od namjernih napada usmjerenih na uklanjanje ili promjenu vodenog žiga. Tehnike za digitalni audio vodeni žig često uključuju manipulaciju audio signala u frekvencijskoj domeni ili korištenje psiho akustičnih modela za skrivanje vodenog žiga u dijelovima signala gdje je najmanje vjerojatno da će ga ljudsko uho percipirati. Zbog sofisticiranosti koja je potrebna u ovim tehnikama, digitalni audio vodeni žig je aktivno područje istraživanja u poljima obrade signala i informacijske sigurnosti.

```
import numpy as np
import scipy.io.wavfile as wav
from scipy import signal

# ucitavanje audio zapisa
stopa, audio = wav.read('SplicingOrgExample.wav')


# kreiranje jednostavnog vodenog žiga tako što dodamo sinusni val na odredjenoj
    frekvenciji
watermark_freq = 1000  # 1 kHz

watermark = np.sin(2 * np.pi * watermark_freq * np.arange(len(audio)) / stopa)

# Dodavanje vodenog žiga u audio zapis
watermarked_audio = audio + watermark


# Spremanje u bazu audio zapisa sa vodenim žigom
wav.write('watermarked.wav', stopa, watermarked_audio)


# simulirajmo scenarij u kojem primamo audio datoteku i želimo provjeriti je li
    uredjena
# Pretpostavit cemo da će svako uredjivanje ukloniti ili izmijeniti vodeni žig


# Ucitavanje potencijalno manipuliranog audio zapisa
stopa_manipulacije, manipuliran_audio = wav.read('CopyMoveForgery.wav')


# Racunanje unakrsne korelacije izmedju uredjenog zvuka i vodenog žiga
cross_correlation = signal.correlate(manipuliran_audio, watermark, mode='same')
```

```
# Ako zvuk nije uredjivan, unakrsna korelacija trebala bi imati vrhunac na
    frekvenciji vodenog žiga

# Koristit ćemo FFT da pronadjemo frekvenciju s ćnajveom magnitudom u unakrsnoj
    korelaciji
freqs = np.fft.rfftfreq(len(cross_correlation), 1 / stopa)
fft = np.abs(np.fft.rfft(cross_correlation))


# đPronaite frekvenciju s maksimalnom čveliinom u FFT-u
max_freq = freqs[np.argmax(fft)]


# Ako je maksimalna frekvencija jednaka frekvenciji vodenog žiga, zvuk nije
    uredjivan
if np.isclose(max_freq, watermark_freq):

    print("Audio␣zapis␣nije␣manipuliran.")

else:

    print("Audio␣zapis␣je␣manipuliran.")
```

Ukratko ću opisati vlastiti algoritam za prepoznavanje krivotvorina metodom vodenog žiga (*eng. "Digital watermarking"*). Verzija digitalno žiga je doista pojednostavljena, koristi 2 audio zapisa, originalni i krivotvoreni. Prilikom kreiranja stavlja se hipoteza da je originalni audio zapis s vodenim žigom spremljen u bazu, odakle dohvaća originalni zapis s vodenim žigom. Simulirajmo scenarij u kojem primamo audio datoteku koju želimo provjeriti je li krivotvorena. Zaboravio sam napomenuti da se ovaj algoritam bazira na kopiraj, premjesti krivotvorinama (*eng. "Copy-move"*). Bilo kakva promjena, manipulacija na audio zapisu će promijeniti ili ukloniti vodeni žig. Za uspostavljanje manipulacije računa se unakrsna korelacija između manipuliranog zvuka i vodenog žiga. Ako zvuk nije uređivan, unakrsna korelacija trebala bi imati vrhunac na frekvenciji vodenog žiga. Za pronalaženje frekvencije s najvećom magnitudom koristimo Brza Fourierova transformacija (*eng. "The Fast Fourier Transform "* FFT). Uzima složeni audio signal koji se mijenja tijekom vremena i rastavlja ga na jednostavne, pojedinačne sinus komponente. Svaki od ovih sinusnih valova ima određenu frekvenciju, amplitudu i fazu. Svaka frekvencija je predstavljena točkom u frekvencijskom domenu. Amplituda je predstavljena magnitudom točke, a faza je predstavljena kutom točke. Pomoću FFT-a pronalazimo frekvenciju s najvećom magnitudom u unakrsnoj koaliciji. Ako je maksimalna frekvencija jednaka frekvenciji vodenog žiga, zvuk nije uređivan.

4.2.2. Spektrogram analiza

Nije za povjerovati koliko mnogo informacija se može izvući iz jednog spektrograma kada je u pitanju digitalni signal. Spektrogram je vizualni prikaz spektra frekvencija u zvuku ili drugom signalu koji se mijenjaju s vremenom. Spektrogrami se intenzivno koriste u područjima glazbe, lingvistike, radara i obrade govora među ostalim. Kroz ovaj rad sam tek naučio važnost spektrograma i načine obrađivanja digitalnog signala. Prije ovog rada nisam niti bio svjestan da postoje ovakve stvari, djelovalo je više kao apstrakcija. Ukratko ću se osvrnuti na metodu spektrogram analize koja obrađuje i filtrira podatke na temelju značajki samog signala. Analizom spektrograma u glasovnim snimkama možemo identificirati sve anomalije ili nedosljednosti koje mogu ukazivati na neovlašteno mijenjanje. Na primjer, možemo tražiti promjene u spektralnoj gustoći, promjene u distribuciji frekvencija ili diskontinuitete u spektrogramu. [54]

Za ovaj praktični primjer koristio sam 4 različite vrste spektrograma u nadi da ću analizom istih uspjeti detektirati krivotvorinu ili manipulaciju iako bi to bilo neostvarivo na realnom slučaju, jer nemaš niti originalni zapis a niti krivotvoreni tj. to su nepoznanice s kojima računalo puno bolje može baratati. U nastavku ću definirate 4 spektrograma koja koristim u svojoj aplikaciji za usporedbu originalnog i krivotvorenog glasovnog zapisa. Ova ideja mi čini zanimljivom jer se nadam pronalasku spoznaje da se ipak može na ovaj način nekako predočiti da se radi o krivotvorini ili deepfake-u. Spektrogrami su: (*eng. "Short-Time Fourier Transform"* STFT), Mel Frequency Cepstral Coefficients (*eng. "Mel Frequency Cepstral Coefficients"* MFCC), Discrete Fourier Transform (*eng. "Discrete Fourier Transform"* DFT) i kromatske značajke (*eng. "Chroma feature"*).

DFT spektrogram je vizualni prikaz spektra frekvencija u diskretnom signalu koji variraju tijekom vremena. Na ovoj vrsti spektrograma X-os predstavlja vrijeme, Y-os predstavlja frekvenciju, a boja ili intenzitet predstavlja veličinu određene frekvencije u određenom trenutku. Međutim, za razliku od STFT spektrograma, DFT spektrogram pretpostavlja da je signal periodičan i ne pruža vremenski razlučivu analizu frekvencije, što ga čini manje prikladnim za nestacionarne signale. DFT spektrogrami korisni su u analizi stacionarnih signala gdje se frekvencijske komponente ne mijenjaju tijekom vremena. [54]

MFCC spektrogrami pružaju vizualni prikaz spektra snage audio signala tijekom vremena, gdje je frekvencijska ljestvica iskrivljena kako bi oponašala ljudsku slušnu percepciju. X-os predstavlja vrijeme, Y-os predstavlja MFCC (svaki koeficijent je značajka koja odgovara različitoj kepstralnoj stopi), a boja ili intenzitet predstavlja veličinu svakog koeficijenta u određenom trenutku. Posebno su korisni u aplikacijama kao što je prepoznavanje govora ili klasifikacija glazbenih žanrova, budući da mogu uhvatiti fonetski važne karakteristike govora. MFCC spektrogrami omogućuju robusnu analizu audio signala, ističući ključne značajke koje nije lako uočiti samo u prikazu vremena ili frekvencije.[53]

STFT spektrogrami pružaju vizualni način provjere sadržaja frekvencije audio signala tijekom vremena. Na ovim dijagramima X-os predstavlja vrijeme, Y-os predstavlja frekvenciju, a boja ili intenzitet označava amplitudu određene frekvencije u određeno vrijeme. Ova metoda omogućuje analizu nestacionarnih signala, otkrivajući koliko različite frekvencije doprinose audio signalu u svakoj točki u vremenu. STFT spektrogrami naširoko se koriste u glazbi,

analizi govora i drugim zadacima obrade zvuka za razumijevanje promjenjivih karakteristika frekvencije audio signala. [54]

Kromatske značajke koje se jednostavno nazivaju (*eng. "Chroma"*), pružaju vizualni prikaz distribucije energije kroz različite visine ili glazbene note u audio signalu tijekom vremena. X-os predstavlja vrijeme, Y-os predstavlja 12 različitih klasa tona (od C do B), a boja ili intenzitet predstavlja razinu energije ili aktivnosti u svakoj klasi tona u određeno vrijeme. Kromagrami su korisni u zadacima pronalaženja glazbenih informacija, kao što su detekcija akorda i procjena ključa, jer hvataju harmonijske i melodijske karakteristike glazbe. Smanjivanjem audio signala na mali skup značajki boje, ovi spektrogrami omogućuju analizu koja je otporna na promjene u tonu i instrumentaciji. [55]

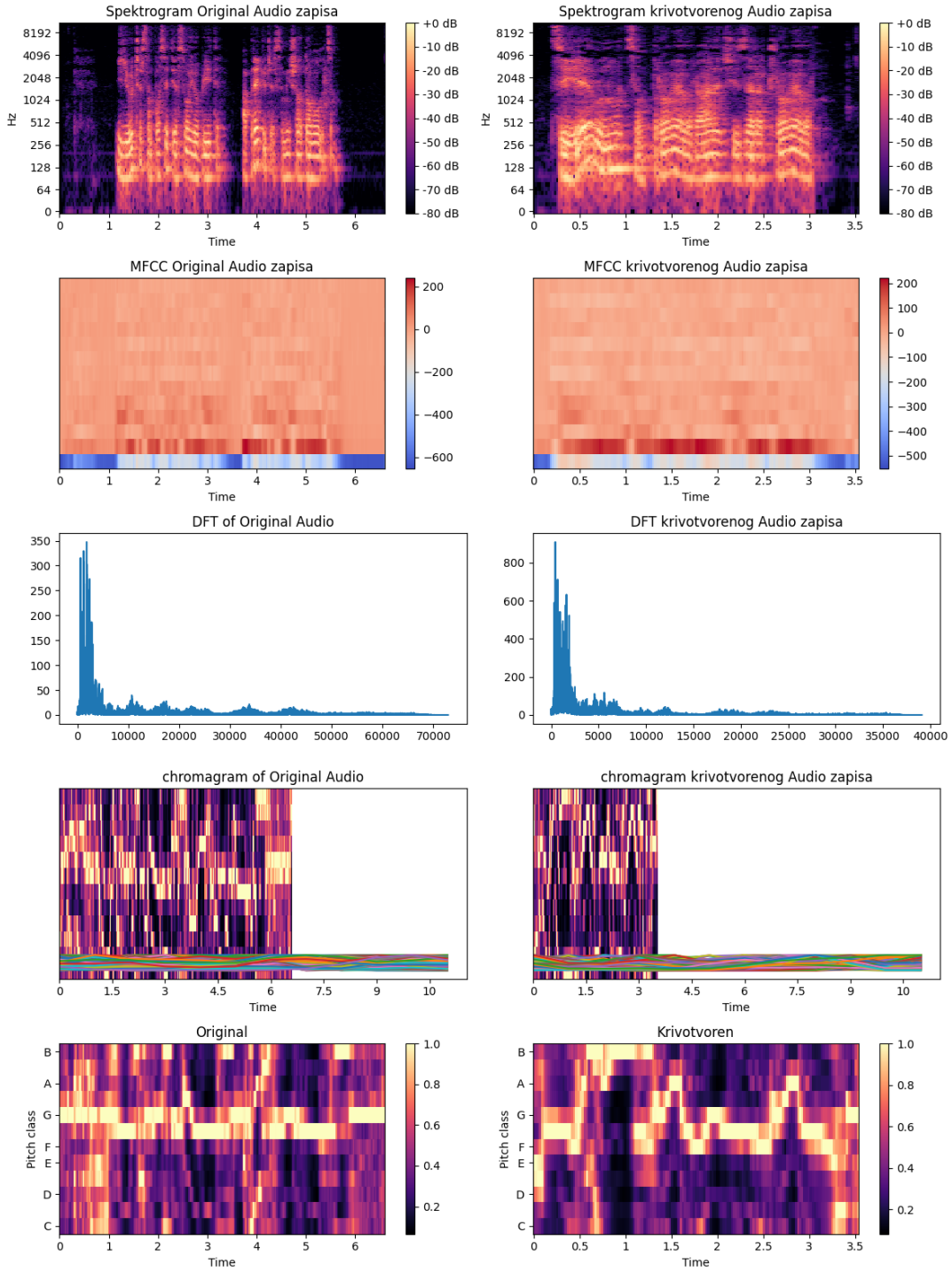
```
import librosa
import librosa.display
import matplotlib.pyplot as plt
import numpy as np
from scipy.fft import fft

# čUitavanje audio zapisa
original_audio_zapis = 'CopyMoveOrgExample.wav'
krivotvoren_audio_zapis = 'CopyMoveForgery.wav'

# Uz ćpomo librose se čuitava audio
original_audio, original_stopa = librosa.load(original_audio_zapis)
krivotvoren_audio, neovlastena_stopa = librosa.load(krivotvoren_audio_zapis)

# čRaunanje the Short-time Fourier Transformacije (STFT)
D_original = librosa.stft(original_audio)
D_krivotvoren = librosa.stft(krivotvoren_audio)

# Konvertiranje vrijednosti signala, amplitude u decibele, relativna vrijednost u
    maksimum
D_original_db = librosa.amplitude_to_db(np.abs(D_original), ref=np.max)
D_krivotvoren_db = librosa.amplitude_to_db(np.abs(D_krivotvoren), ref=np.max)

# čRauanje MFCC čznaajki
mfcc_original = librosa.feature.mfcc(y=original_audio, sr=original_stopa, n_mfcc=13)

mfcc_krivotvoren = librosa.feature.mfcc(y=krivotvoren_audio, sr=neovlastena_stopa,
    n_mfcc=13)

# čRaunanje Discrete Fourier transformacije (DFT)
dft_original = np.abs(fft(original_audio)[:len(original_audio)//2])
dft_krivotvoren = np.abs(fft(krivotvoren_audio)[:len(krivotvoren_audio) // 2])

# čRaunaj chroma čznaajke
chroma_original = librosa.feature.chroma_stft(original_audio, sr=original_stopa)

chroma_krivotvoren = librosa.feature.chroma_stft(krivotvoren_audio, sr=
    neovlastena_stopa)
```

```
# Generiranje  chromagrams
C_original = librosa.feature.chroma_cqt(original_audio, sr=original_stopa)
C_krivotvoren = librosa.feature.chroma_cqt(krivotvoren_audio, sr=neovlastena_stopa)


# postavi
plt.figure(figsize=(12, 16))

# Original audio spectrogram
plt.subplot(5, 2, 1)

librosa.display.specshow(D_original_db, sr=original_stopa, x_axis='time', y_axis='
    log')

plt.colorbar(format='%+2.0f␣dB')
plt.title('Spektrogram␣Original␣Audio␣zapisa')

# krivotvorenog audio spectrogram
plt.subplot(5, 2, 2)
librosa.display.specshow(D_krivotvoren_db, sr=neovlastena_stopa, x_axis='time',
    y_axis='log')

plt.colorbar(format='%+2.0f␣dB')

plt.title('Spektrogram␣krivotvorenog␣Audio␣zapisa')

# Original audio MFCC
plt.subplot(5, 2, 3)
librosa.display.specshow(mfcc_original, sr=original_stopa, x_axis='time')

plt.colorbar()
plt.title('MFCC␣Original␣Audio␣zapisa')

# Krivotvorenog audio MFCC
plt.subplot(5, 2, 4)
librosa.display.specshow(mfcc_krivotvoren, sr=neovlastena_stopa, x_axis='time')

plt.colorbar()
plt.title('MFCC␣krivotvorenog␣Audio␣zapisa')

# Original audio DFT
plt.subplot(5, 2, 5)

plt.plot(dft_original)
plt.title('DFT␣of␣Original␣Audio')

# Tampered audio DFT
plt.subplot(5, 2, 6)
plt.plot(dft_krivotvoren)

plt.title('DFT␣krivotvorenog␣Audio␣zapisa')

# Original audio chromagram
```

```
plt.subplot(5, 2, 7)
librosa.display.specshow(chroma_original, sr=original_stopa, x_axis='time')

plt.plot(chroma_original)
plt.title('chromagram␣of␣Original␣Audio')

# Krivotvoren audio chromagram
plt.subplot(5, 2, 8)
librosa.display.specshow(chroma_krivotvoren, sr=neovlastena_stopa, x_axis='time')

plt.plot(chroma_krivotvoren)
plt.title('chromagram␣krivotvorenog␣Audio␣zapisa')

# Prikaz the original chromagram
plt.subplot(5, 2, 9)
librosa.display.specshow(C_original, sr=original_stopa, x_axis='time', y_axis='
    chroma')

plt.title('Original')
plt.colorbar()

# Prikaz  krivotvorenog chromagram
plt.subplot(5, 2, 10)
librosa.display.specshow(C_krivotvoren, sr=neovlastena_stopa, x_axis='time', y_axis=
    'chroma')

plt.title('Krivotvoren')
plt.colorbar()

# Prikaz
plt.tight_layout()
plt.show()
```

4.2.2.1. Scenarij 1

U Scenariju 1 sam kreirao krivotvorinu s kopiraj, premjesti (*eng. "Copy-Move"*) audio zapis, kako bi ga mogao usporediti s originalnim. Cilj ovog scenarija je uočiti manipulaciju na pomoću spektrograma koji su generirani. (slika 33) prikazuje spektrograme STFT, MFCC, DFT i Kroma svojstva su prikazan dva puta. Jedina razlika koja se može uočiti je na STFT spektrogramu gdje se vidi da je audio malo isprekidan, nema harmonije kao što je kod originalnog zapisa. Po mojem skromnom iskustvu ovaj zapis bi definitivno išao na detaljniju analizu. Ako se prisjetimo kada smo kreirali kopiraj, premjesti (*eng. "Copy-Move"*) krivotvorinu. Kontekst je glasio "Jučer sam posjetio svoju obitelj, a prekjučer sam išao tom ulicom". Kao mali podsjetnik što smo odradili prije, zamijenili smo riječi "Jučer" i "prekjučer".

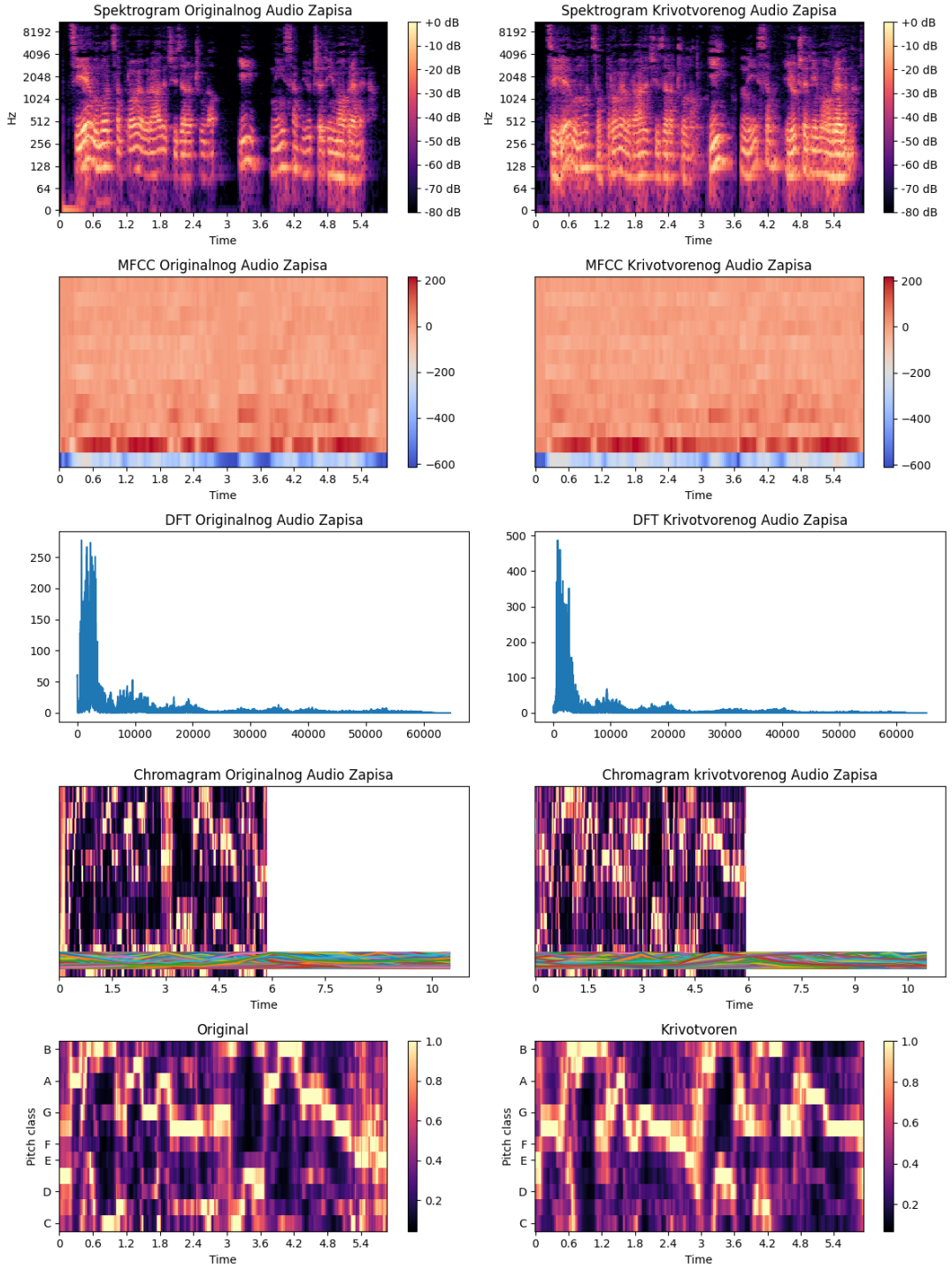


Slika 33: Usporedba originalnog i krivotvorenog glasovnog zapisa (*eng. "Copy-Move"*)

4.2.2.2. Scenarij 2

U Scenariju 2 smo uzeli krivotvorinu spajanja (*eng. "Splicing"*) i usporedili smo ju s novo kreiranim glasovnim zapisom. Nastojao sam da bude sličan kao krivotvorenom pa pauzama i naglašavanju kako se razlike ne bi vidjele očigledno. No u ovom primjeru (slika 34) se doista ništa ne može primijetiti što bi naslutilo da je jedno krivotvorina a drugo nije. Poneke stvari izgledaju slično iako je desni sastavljen od 3 druga audio zapisa. Postoji povezanost jer se

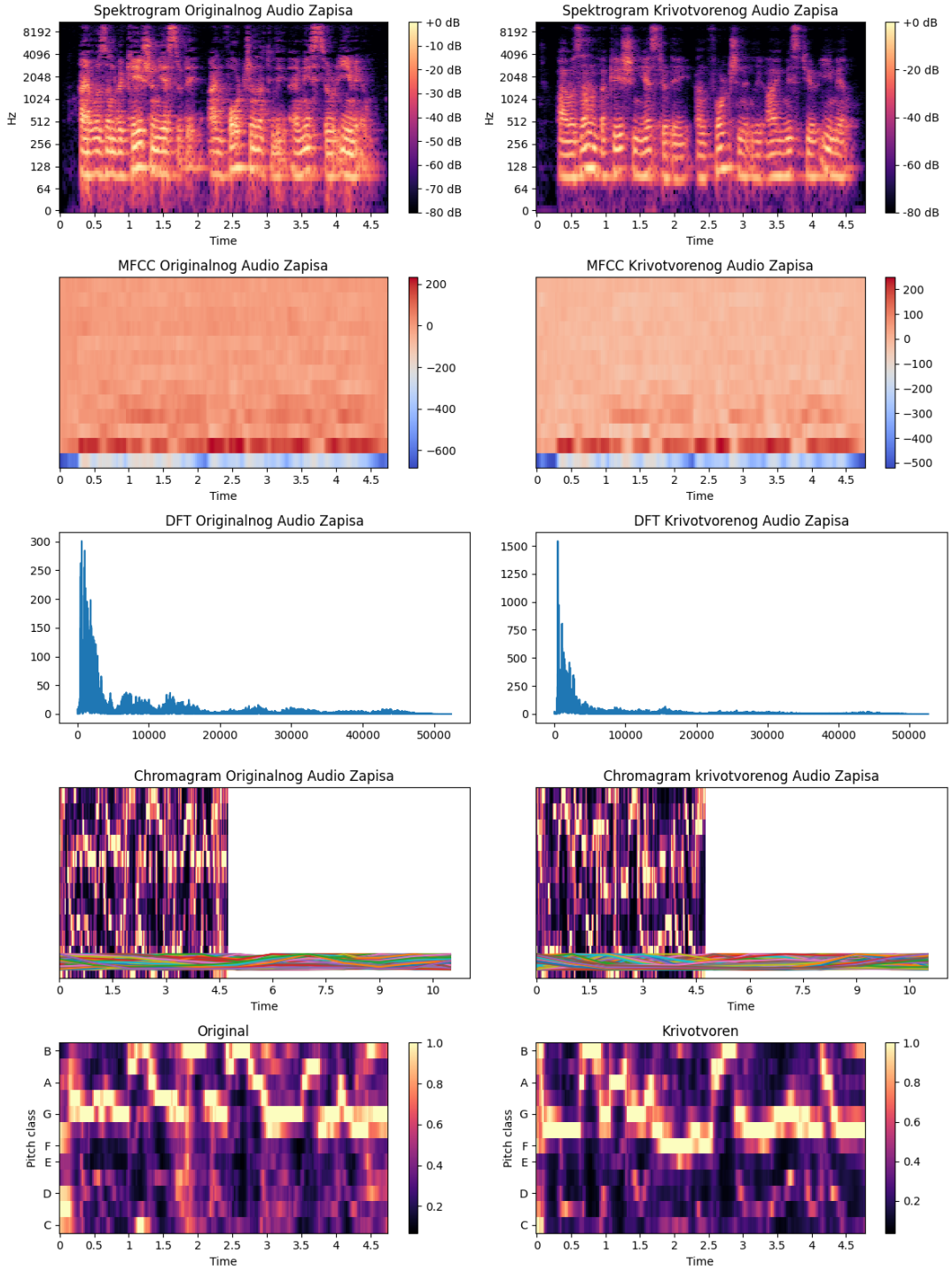
ovdje radi o kontekstu "Cijelu noć sam proveo radeći na računalu i nisam jučer imao vremena". Ovo "i" se jako ističe pogotovo u originalnom STFT spektrogramu ako malo bolje pogledate, čak se i kod krivotvorenog može uočiti. Na temelju STFT možda bi se dalo naslutiti da je krivotvorina ali mislim ipak da je to stručne ljude koji to znaju iščitavati.



Slika 34: Usporedba originalnog i krivotvorenog glasovnog zapisa spajanjem (*eng. "Splicing"*)

4.2.2.3. Scenarij 3

U Scenariju 3 uspoređujemo sintetički glas kreiran od Resemble.AI aplikacije i moj vlastiti glas, glasovni zapisi su na engleskom jeziku. Kontekst glasovnog zapisa glasi "Bio sam na odmoru jučer, propustio sam tvoje mailove nažalost" (*eng. "I was on vacation yesterday, unfortunately i missed your email."*). Na temelju spektrograma (slika 35) ovdje čak najviše MFCC spektrogram ukazuje da se radi o blijedom glasu jer ne sadrži emocija ako se usporedi da originalnim koji gdje su boje narančasto-crvene. Prema STFT vidimo identični slučaj kako je originalni glas istaknutiji, življi. Za ovaj primjer bi se moglo zaključiti da se radi o sintezi jer se primijeti po navedenim spektrogramima da nedostaje emocije u glasu.



Slika 35: Usporedba originalnog i krivotvorenog glasovnog zapisa, sinteza govora (*eng. "Speech Synthesis"*)

5. Zaključak

Uistinu ne mogu vjerovati da je kreiranje krivotvorina izuzetno lakše za napraviti nego kreirati metodu koja će prepoznati manipuliran glasovni zapis. Iznenađen sam koliko različitih krivotvorina postoji: spajanje, kopiraj, premjesti, sinteza govora, repriza, impersoniranje i druge. Podijelili smo metode za pronalaženje krivotvorina na pasivne i aktivne, iz nekog razloga većina znanstvenika se fokusira na pasivne metode slijepim otkrivanjem (*eng"Blind detection"*) [56] temeljene na strojnom učenju, dubokom učenju ili/i neuronskim mrežama. Uz naveden mnogo članaka sam vidio koje se bave detekcijom audioSpoof krivotvorina te da postoje. Kamble, Sailor, Patil i dr. je nabrojao da od 2015. godine održavaju sastanci i natjecanja ASVSpoof, koje se svake druge godine ponovi, kojem je cilj skupiti podatke za testiranje metoda detekcija protiv deepfake krivotvorina i sinteze govora. U posljednjem dijelu je definirana nekolicina metoda s kojima se mogu detektirati krivotvorine ali svaka metoda nosi nešto svoje i podređena je jednoj vrsti krivotvorina. Nisam pronašao univerzalnu metodu koja može svaku detektirati. Smatram da su najefikasnije metode koje koriste neuronske mreže i duboko učenje ali su i najkompleksnije za napraviti. U praktičnom radu smo vidjeli i pokazali da se jako teško mogu krivotvorine detektirati samo pomoću analizom spektrograma, ali vjerujem kad bi se uzelo malo vremena i produbilo znanje za isto. Moglo bi se puno više stvari zamijetiti, očaran sam spektrogramima i jednostavnost kreiranja istih. Kako sam već napomenuo malo sam iznenađen da još nema globalnih algoritama za pronalaženje krivotvorina. I općenito dosta nove literature nije za širu publiku i jako malo ljudi iz Europe se bavi audio forenzikom. Većina autora je iz Azije što mi je bilo fascinantno za vidjeti, jer sam imao osjećaj da se Europa zalaže za globalnu sigurnost.

Popis literature

[1] M. Stikic, M. Vujovic i A. Kartelj, „A comparative study of different approaches for solving the traveling salesman problem", *Proceedings of the Conference on Business Information Systems and E-Business Technologies*, 2014., str. 59–68.

[2] A. Trade, *Neue Betrugsmasche: Fake President mit Stimmimitation*, 2022. adresa: https://www.allianz-trade.de/content/dam/onemarketing/aztrade/allianz-trade_de/presse/neue-betrugsmasche-fake-president-mit-stimmimitation.pdf.

[3] M. Koenig, *History of Audio Forensics*, 2019. adresa: https://www.mediamedic.studio/history-of-audio-forensics/.

[4] R. C. Maher, „History of Audio Forensics", *Principles of Forensic Audio Analysis*. Cham: Springer International Publishing, 2018., str. 29–37, ISBN: 978-3-319-99453-6. DOI: 10.1007/978-3-319-99453-6_3. adresa: https://doi.org/10.1007/978-3-319-99453-6_3.

[5] G. Van Rossum i dr., „Python Programming Language.", *USENIX annual technical conference*, Santa Clara, CA, sv. 41, 2007., str. 1–36.

[6] Resemble.ai, *Resemble.ai*, 2023. adresa: https://www.resemble.ai/ (pogledano 16. 4. 2023.).

[7] A. Durbin, *The AI Team that Brought Back Andy Warhol*, 2019. adresa: https://www.frieze.com/article/ai-team-brought-back-andy-warhol (pogledano 13. 4. 2023.).

[8] L. Tulchak i A. Marchuk, „History of Python", disertacija, 2016.

[9] „TIOBE Index for April 2023", 2023. adresa: https://www.tiobe.com/tiobe-index/ (pogledano 10. 4. 2023.).

[10] *Python Logo*. adresa: https://www.python.org/community/logos/ (pogledano 10. 4. 2023.).

[11] K. Srinath, „Python–the fastest growing programming language", *International Research Journal of Engineering and Technology*, sv. 4, br. 12, str. 354–357, 2017.

[12] Movavi, *What Audio Format Is the Best?*, 2022. adresa: https://www.movavi.io/what-audio-format-is-the-best-2/.

[13] B. McFee, M. McVicar, C. Raffel i dr., *Librosa*, travanj 2023. adresa: https://librosa.org/doc/latest/index.html (pogledano 5. 4. 2023.).

[14] S. Davis i P. Mermelstein, „Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, sv. 28, br. 4, str. 357–366, 1980. DOI: 10.1109/TASSP.1980.1163420.

[15] P. Raguraman, R. Mohan i M. Vijayan, „Librosa based assessment tool for music information retrieval systems”, *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, IEEE, 2019., str. 109–114.

[16] B. McFee, C. Raffel, D. Liang i dr., „librosa: Audio and music signal analysis in python”, *Proceedings of the 14th python in science conference*, sv. 8, 2015., str. 18–25.

[17] F. Lazzeri, *Machine learning for time series forecasting with Python*. John Wiley & Sons, 2020.

[18] Martín Abadi, Ashish Agarwal, Paul Barham i dr., *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, Software available from tensorflow.org, 2015. adresa: https://www.tensorflow.org/.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort i dr., „Scikit-learn: Machine Learning in Python”, *Journal of Machine Learning Research*, sv. 12, br. 85, str. 2825–2830, 2011. adresa: http://jmlr.org/papers/v12/pedregosa11a.html.

[20] F. Chollet, *Deep Learning with Python*. Manning Publications, 2018.

[21] B. Balamurali, K. E. Lin, S. Lui, J.-M. Chen i D. Herremans, „Toward robust audio spoofing detection: A detailed comparison of traditional and learned features”, *IEEE Access*, sv. 7, str. 84 229–84 241, 2019.

[22] J. D. Hunter, „Matplotlib: A 2D Graphics Environment”, *Computing in Science & Engineering*, sv. 9, br. 3, str. 90–95, 2007. DOI: 10.1109/MCSE.2007.55.

[23] Z. Ali, M. Imran i M. Alsulaiman, „An Automatic Digital Audio Authentication/Forensics System”, *IEEE Access*, sv. PP, str. 1–1, veljača 2017. DOI: 10.1109/ACCESS.2017.2672681.

[24] A. Team, *Audacity® software is copyright © 1999-2021*, 2023. adresa: https://audacityteam.org/ (pogledano 16. 4. 2023.).

[25] M. A. Qamhan, H. Altaheri, A. H. Meftah, G. Muhammad i Y. A. Alotaibi, „Digital Audio Forensics: Microphone and Environment Classification Using Deep Learning”, *IEEE Access*, sv. 9, str. 62 719–62 733, 2021. DOI: 10.1109/ACCESS.2021.3073786.

[26] M. Zakariah, M. Khan i H. Malik, „Digital multimedia audio forensics: past, present and future”, *Multimedia Tools and Applications*, sv. 77, 2018. DOI: 10.1007/s11042-016-4277-2.

[27] P. R. Bevinamarad i M. Shirldonkar, „Audio Forgery Detection Techniques: Present and Past Review”, *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, 2020., str. 613–618. DOI: 10.1109/ICOEI48184.2020.9143014.

[28] C. Zeng, D. Zhu, Z. Wang, Z. Wang, N. Zhao i L. He, „An end-to-end deep source recording device identification system for Web media forensics”, *International Journal of Web Information Systems*, sv. 16, br. 4, str. 413–425, 2020. DOI: 10.1108/IJWIS-06-2020-0038. adresa: https://doi.org/10.1108/IJWIS-06-2020-0038 (pogledano 22. 3. 2023.).

[29] Z. Khanjani, G. Watson i V. P. Janeja, „Audio deepfakes: A survey”, en, *Front. Big Data*, sv. 5, str. 1 001 063, 2022.

[30] V. Rahinj, R. Patole i S. Metkar, „Active Learning Based Audio Tampering Detection”, *2022 International Conference on Connected Systems & Intelligence (CSI)*, 2022., str. 1–5. DOI: 10.1109/CSI54720.2022.9923997.

[31] B. Ustubioglu, G. Tahaoglu i G. Ulutas, „Detection of audio copy-move-forgery with novel feature matching on Mel spectrogram”, *Expert Systems with Applications*, sv. 213, str. 118 963, 2023., ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.118963. adresa: https://www.sciencedirect.com/science/article/pii/S0957417422019819.

[32] Z. Su, M. Li, G. Zhang i dr., „Robust Audio Copy-Move Forgery Detection Using Constant Q Spectral Sketches and GA-SVM”, *IEEE Transactions on Dependable and Secure Computing*, str. 1–15, 2022. DOI: 10.1109/TDSC.2022.3215280.

[33] Z. Ali, M. Imran i M. Alsulaiman, „An Automatic Digital Audio Authentication/Forensics System”, *IEEE Access*, sv. 5, str. 2994–3007, 2017. DOI: 10.1109/ACCESS.2017.2672681.

[34] X. Meng, C. Li i L. Tian, „Detecting Audio Splicing Forgery Algorithm Based on Local Noise Level Estimation”, *2018 5th International Conference on Systems and Informatics (ICSAI)*, 2018., str. 861–865. DOI: 10.1109/ICSAI.2018.8599318.

[35] Z. Mubeen, M. Afzal, Z. Ali, S. Khan i M. Imran, „Detection of impostor and tampered segments in audio by using an intelligent system”, *Computers & Electrical Engineering*, sv. 91, str. 107 122, 2021., ISSN: 0045-7906. DOI: https://doi.org/10.1016/j.compeleceng.2021.107122. adresa: https://www.sciencedirect.com/science/article/pii/S0045790621001269.

[36] Z. Almutairi i H. Elgibreen, „A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions”, *Algorithms*, sv. 15, br. 5, 2022., ISSN: 1999-4893. DOI: 10.3390/a15050155. adresa: https://www.mdpi.com/1999-4893/15/5/155.

[37] M. R. Kamble, H. B. Sailor, H. A. Patil i H. Li, „Advances in anti-spoofing: from the perspective of ASVspoof challenges”, *APSIPA Transactions on Signal and Information Processing*, sv. 9, e2, 2020. DOI: 10.1017/ATSIP.2019.21.

[38] M. Sahidullah, H. Delgado, M. Todisco i dr., „Introduction to Voice Presentation Attack Detection and Recent Advances”, *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, S. Marcel, J. Fierrez i N. Evans, ur. Singapore: Springer Nature Singapore, 2023., str. 339–385, ISBN: 978-981-19-5288-3. DOI: 10.1007/978-981-19-5288-3_13. adresa: https://doi.org/10.1007/978-981-19-5288-3_13.

[39] S. Gupta, S. Cho i C.-C. J. Kuo, „Current Developments and Future Trends in Audio Authentication”, *IEEE MultiMedia*, sv. 19, br. 1, str. 50–59, 2012. DOI: 10.1109/MMUL.2011.74.

[40] Z. Wang, Y. Yang, C. Zeng, S. Kong, S. Feng i N. Zhao, „Shallow and deep feature fusion for digital audio tampering detection”, *EURASIP Journal on Advances in Signal Processing*, sv. 2022, br. 1, str. 69, kolovoz 2022. DOI: 10.1186/s13634-022-00900-4.

[41] G. K. Birajdar i V. H. Mankar, „Digital image forgery detection using passive techniques: A survey”, *Digital Investigation*, sv. 10, br. 3, str. 226–245, 2013., ISSN: 1742-2876. DOI: https://doi.org/10.1016/j.diin.2013.04.007. adresa: https://www.academia.edu/31977007/Digital_image_forgery_detection_using_passive_techniques_A_survey.

[42] K. Muroi, K. Kondo i S. Takahashi, „Speech Manipulation Detection Method using Audio Watermarking”, *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, 2021., str. 7–8. DOI: 10.1109/GCCE53005.2021.9621898.

[43] T. Arbi, B. Geller i O. P. Pasquero, „Direct-Sequence Spread Spectrum with Signal Space Diversity for High Resistance to Jamming”, *MILCOM 2021 - 2021 IEEE Military Communications Conference (MILCOM)*, 2021., str. 670–676. DOI: 10.1109/MILCOM52596.2021.9652967.

[44] M. Malekesmaeili i R. K. Ward, „A local fingerprinting approach for audio copy detection”, *Signal Processing*, sv. 98, str. 308–321, 2014., ISSN: 0165-1684. DOI: https://doi.org/10.1016/j.sigpro.2013.11.023. adresa: https://www.sciencedirect.com/science/article/pii/S0165168413004593.

[45] M. R. R. Ansori, Allwinnaldo, R. N. Alief, I. S. Igboanusi, J. M. Lee i D.-S. Kim, „HADES: Hash-based Audio Copy Detection System for Copyright Protection in Decentralized Music Sharing”, *IEEE Transactions on Network and Service Management*, str. 1–1, 2023. DOI: 10.1109/TNSM.2023.3241610.

[46] G. Ulutas, A. Ustubioglu, B. Ustubioglu i G. Tahaoglu, „Localization of Forgery on Audio Clips Using GLCM Features and Mel Spectograms”, *2022 45th International Conference on Telecommunications and Signal Processing (TSP)*, 2022., str. 314–317. DOI: 10.1109/TSP55681.2022.9851294.

[47] H. Malik i H. Farid, „Audio forensics from acoustic reverberation”, *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010., str. 1710–1713. DOI: 10.1109/ICASSP.2010.5495479.

[48] W. Qiu, W. J. Murphy i A. Suter, „Kurtosis: A new tool for noise analysis”, *Acoust Today*, sv. 16, br. 4, str. 39–47, 2020.

[49] Garofolo, John S., Lamel, Lori F., Fisher, William M. i dr., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, 1993. DOI: 10.35111/17GK-BN40. adresa: https://catalog.ldc.upenn.edu/LDC93S1.

[50] X. Pan, X. Zhang i S. Lyu, „Detecting splicing in digital audios using local noise level estimation”, *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012., str. 1841–1844. DOI: 10.1109/ICASSP.2012.6288260.

[51] A. Vaswani, N. Shazeer, N. Parmar i dr., *Attention Is All You Need*, 2017. arXiv: 1706.03762 [cs.CL].

[52] E. A. AlBadawy, S. Lyu i H. Farid, „Detecting AI-Synthesized Speech Using Bispectral Analysis", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, lipanj 2019.

[53] A. K. Singh i P. Singh, „Detection of AI-Synthesized Speech Using Cepstral & Bispectral Statistics", *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2021., str. 412–417. DOI: 10.1109/MIPR51284.2021.00076.

[54] J. O. Smith", *Spectral Audio Signal Processing*. ccrma.stanford.edu, 2011.

[55] G. K. Birajdar i M. D. Patil, „Speech/music classification using visual and spectral chromagram features", *Journal of Ambient Intelligence and Humanized Computing*, sv. 11, br. 1, str. 329–347, travanj 2019. DOI: 10.1007/s12652-019-01303-4. adresa: https://doi.org/10.1007/s12652-019-01303-4.

[56] M. Imran, Z. Ali, S. T. Bakhsh i S. Akram, „Blind Detection of Copy-Move Forgery in Digital Audio Forensics", *IEEE Access*, sv. 5, str. 12 843–12 855, 2017. DOI: 10.1109/ACCESS.2017.2717842.

Popis slika

Popis tablica

1. Prilog 1

2. Prilog 2