Elektronski fakultet u Nišu

Augmentacija audio podataka u mašinskom učenju

Student: Danilo Milošević

Profesor: Aleksandar Stanimirović

Sadržaj

1. Uvod

U toku poslednje decenije smo svedoci sve boljih i naprednijih modela veštačke inteligencije, kako kod obrade slika, teksta, videa tako i kod audio podataka. Kao i kod drugih oblasti mašinskog učenja, performanse modela zavise dosta od kvaliteta i količine podataka nad kojima se trenira model. Prikupljanje velike količine kvalitetnih audio podataka (pogotovo anotiranih) je dosta skupo, vremenski zahtevno a u nekim slučajevima i nemoguće. Zbog ovoga primenjujemo augmentaciju podataka.

Augmentacija audio podataka, i uopšte bilo koja vrsta augmentacije podataka, predstavlja skup tehnika kojima se od postojećeg skupa podataka generišu dodatni podaci primenom različitih transformacija. Za razliku od augmentacije slika, koja je danas standardna praksa, augmentacija audio podataka ima dodatne izazove zbog vremenske prirode audio signala, kao i kompleksnosti percepcije zvuka kod ljudi.

Osnovni cilj augmentacije je da se poveća raznovrsnost i veličina trening skupa bez sakupljanja novih podataka, čime se model čini otpornijim na varijacije koje se mogu desiti tokom primene modela (šum npr). Jedan primer je prepoznavanje ljudskog govora. Ukoliko je model treniran na ljudskim glasovima snimljenim u studiju, postoji mogućnost da će performanse modela biti loše u realnim uslovima - na ulici, u kafiću, kod kuće itd.

U ovom radu ćemo prvenstveno obraditi šta je zvuk, kako se može opisati i predstaviti, kao i kakve različite tehnike augmentacije podataka. Razmotrićemo kako jednostavnije tako i kompleksnije transformacije, u različitim domenima, kada i kako treba primeniti date tehnike kao i njihove prednosti i nedostatke. Na kraju ćemo spomenuti kako možemo meriti efikasnost tehnika augmentacije koje će kasnije biti obrađene u praktičnom delu projekta.

2. Osnove audio signala

2.1 Digitalni audio signal

Zvuk je mehanički talas koji se prostire kroz neku sredinu kao rezultat vibracija. Kako bi se zvuk sačuvao na računaru on se zapisuje u digitalnom obliku, gde je predstavljen diskretizovanom verzijom kontinualnog zvučnog talasa. Proces digitalizacije se sastoji od semplovanja (sampling) i kvantizacije.

- **Sampling**
  - Proces merenja amplitude zvučnog talasa u regularnim vremenskim intervalima. Frekvencija *sampling*-a određuje koliko puta u sekundi ćemo meriti amplitudu zvučnog signala. Po Nyquist-Shannon-ovoj teoremi semplovanja signala, potrebno je semplovati sa bar duplo manjom frekvencijom od izvorne frekvencije signala. U suprotnom dolazi do *aliasing*a

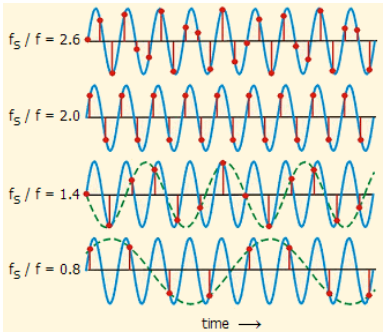


Figure 1: Semplovanje audio signala

Standardne frekvencije semplovanja (kHz)	Upotreba
8	telefonski govor
44.1	CD kvalitet
48	profesionalni audio
96 ili 192	high-res audio

Table 1: Broj vrednosti amplituda u zavisnosti od broja bitova

- **Kvantizacija**
  - Kako su računari digitalni uređaji konačne preciznosti, potrebno je vrednosti amplitude signala mapirati na jednu od 2^n vrednosti, pri čemu n predstavlja broj bitova kojim možemo predstaviti vrednosti amplitude.

Broj bitova	Broj mogućih vrednosti amplitude
8	256
16	65,536 (CD)
24	16,777,216 (profesionalni audio)
32	4,294,967,296 (digitalna obrada signala)

Table 2: Broj vrednosti amplituda u zavisnosti od broja bitova

Audio signal se zatim predstavlja kao diskretan niz vrednosti $x[n]$. Kod mono signala to je jednodimenzionalni niz, dok stereo signal ima dva kanala.

2.2 Reprezentacije audio signala

Audio signali se mogu predstaviti i analizirati na više načina, pri čemu se uglavnom predstavljaju u vremenskom ili u frekventnom domenu.

- **Vremenski domen**

  Najjednostavnija reprezentacija audio signala - prikaz amplitude signala kroz vreme.

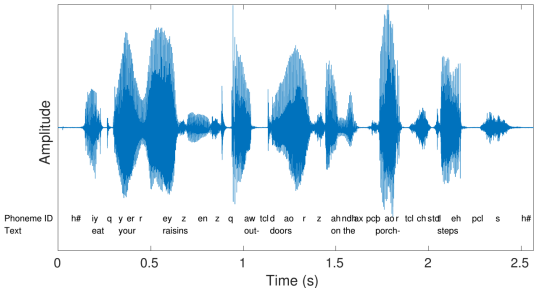


Figure 2: Prikaz audio signala u vremenskom domenu

Prednosti	Nedostaci
Vremenske informacije su očuvane	Nepogodno za analizu frekvencija
Nema gubitaka tokom konverzije	Zahteva duže sekvence za neuronske mreže
Lako za razumevanje i prikaz	

Table 3: Prednosti i nedostaci prikaza audio signala u vremenskom domenu

- **Frekventni domen**

  Furijeovom transformacijom je moguće izvršiti dekompoziciju audio signala na njegove frekvencije. Analiza signala korišćenjem Furijeove transformacije je pogodna za slučajeve kada frekvencije signala ne variraju vremenom. U realnosti je ovo retkost, s obzirom da ljudski govor i muzika variraju vremenom, tj oni su aperiodični signali.

  Zato umesto direktnog vršenja Furijeove transformacije nad signalom, možemo ulazni signal da podelimo na preklapajuće segmente a zatim nad njima vršiti FFT. Podelom audio signala na kratke preklapajuće segmente, a zatim primenom Furijeove transformacije dobijamo STFT - *Short Time Fourier Transform*.

  Matematički STFT se može definisati kao

$$\mathrm{STFT}\{x[n]\}(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]\; w[n-m]\; e^{-j\omega n} \tag{1}$$

  pri čemu je $x[n]$ ulazni signal u trenutku n, dok je w window funkcija i m vreme u okviru kog analiziramo frekvence.

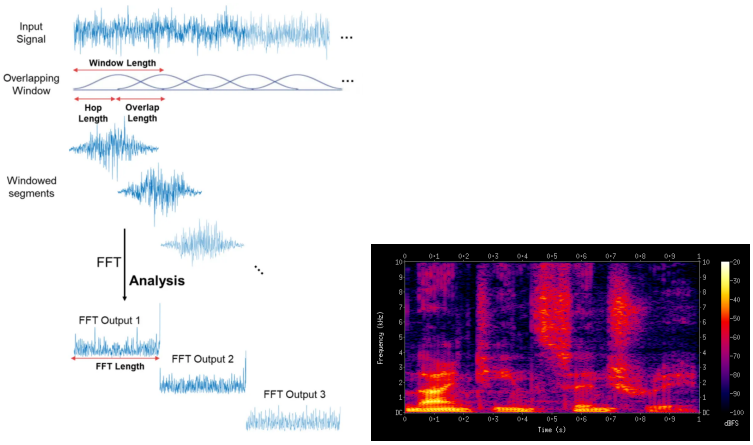


Figure 3: Levo - Proces dobijanja STFT reprezentacije. Desno - STFT reprezentacija. X osa predstavlja vreme, Y osa frekvence a boja intenzitet.

Postoji više window funkcija, ispod je dat primer Hann-ove window funkcije

$$w[n] = \begin{cases} \frac{1}{2}\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right), & 0 \leq n < N, \\ 0, & \text{u suprotnom.} \end{cases} \tag{2}$$

STFT reprezentacija međutim ne opisuje najbolje kako ljudsko uho percipira zvuk. Na osnovu istraživanja je zaključeno da ljudi najbolje registruju i razlikuju zvukove u rasponu frekvenci od 500 do 1000Hz, dok teško razlikuju više frekvence npr 1-1.5kHz. Upravo zato je osmišljena mel skala. Mel skala se dobija pretvaranjem linearnog spektrograma u logaritamski na sledeći način

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right)$$

pri čemu je m rezultujuća vrednost a f ulazna frekvenca.

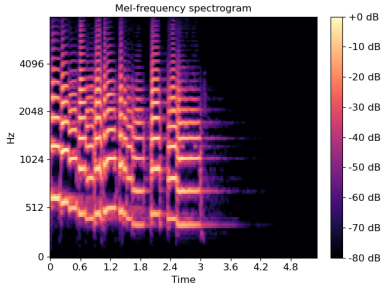


Figure 4: Mel spektrogram

Mel spektrogram je pogotovo koristan kod zadataka prepoznavanja govora i muzike.

2.3 Izazovi u radu sa audio podacima

Rad sa audio podacima kod mašinskog učenja donosi nekoliko specifičnih izazova

- **Promenljivo trajanje zvučnih zapisa**

  Za razliku od slika koje imaju fiksne dimenzije, audio snimci mogu biti proizvoljnog trajanja. Model mora biti u stanju da procesira signale različite dužine što može da uradi

  - Korekcijom dužine snimka, skraćivanjem ili padd-ovanjem signala
  - Korišćenjem arhitektura koje primaju ulaze različitih dužina (rekurentne mreže, transformeri)

- **Velika dimenzionalnost**

  Audio sa frekvencijom semplovanja od 16kHz ima 16000 vrednosti u sekundi. Za snimak od 10 sekundi, to je 160000 vrednosti. Zbog toga se često koriste kompresovane reprezentacije.

- **Osetljivost na šum**

  Zvukovi snimljeni u realnim uslovima često sadrže različite vrste šuma kao što su pozadinski šum, električni šum mikrofona, odjek prostorija, zvukove iz više različitih izvora koji se prepliću (govor više ljudi u isto vreme) itd.

- **Nedostatak label-ovanih podataka**

  Kvalitetni, labelovani audio skupovi podataka su relativno mali i skupi. Problem je u tome što je za govor potrebno ručno izvršiti transkripciju sa preciznom vremenskom anotacijom, dok je za muziku potrebna kategorizacija žanrova, instrumenata, takta itd.

- **Varijabilnost govora**

  Kod prepoznavanja govora mnogo faktora utiče na audio signal - jezik, akcenat, pol, emotivno stanje, brzina govora, artikulacija, pozadinski šum itd.

Baš zbog ovih izazova je augmentacija audio podataka ključna za treniranje modela koji će imati dobre performanse u realnim uslovima.

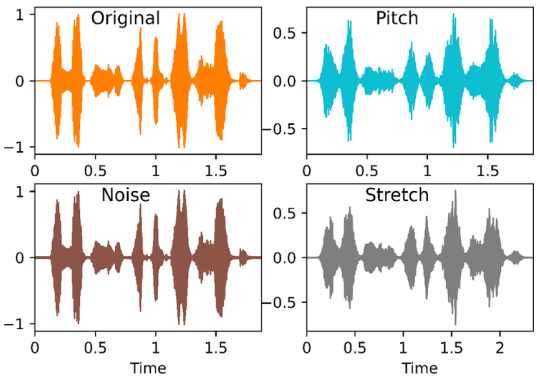


Figure 5: Nekoliko vrsta augmentacija audio signala

3. Osnove augmentacije audio signala

Augmentacija podataka je tehnika u mašinskom učenju koja uključuje kreiranje novih trening primera primenom transformacija na postojeće podatke. Metode augmentacije podataka su dizajnirane tako da uvedu dovoljno varijacije u skupu trening podataka tako da održe realističnost i semantički sadržaj ali i poboljšaju sposobnost modela da generalizuje.

Postoji više podela augmentacija audio podataka

1. Po domenu primen
   - **Augmentacije u vremenskom domenu**
     Primenjuju se direktno na waveform audio zapisa
   - **Augmentacije u frekventnom domenu**
     Primenjuju se na spectrogram audio zapisa
   - **Augmentacije u feature prostoru**
     Primenjuju se na ekstrakovnim feature-ima
2. Po složenosti
   - **Jednostavne transformacije**
     Dodavanje šuma, promena pitch-a
   - **Kompozitne transformacije**
     Nastaju kombinaciju više jednostavnih transformacija
   - **Naučene transformacije**
     Transformacije primenom treniranih modela kao što su GAN ili varijacioni auto-enkoderi
3. Po determinističnosti
   - **Determinističke**
     Ista transformacija daje iste rezultate
   - **Stohastičke**
     Uvode nasumičnost tokom tranformacija

Ciljevi augmentacije podataka su

- **Povećanje veličine trening skupa**
  Osnovni cilj je kreiranje većeg i raznovrsnijeg trening skupa bez skupljanja novih podataka. Duboke neuronske mreže, posebno konvolucione mreže i transformeri, zahtevaju velike količine podataka da bi efikasno naučile reprezentacije podataka.

- **Regularizacija i smanjenje overfitting-a**
  Uvođenjem varijacije u trening podatke, sprečavamo model da "zapamti" specifične detalje trening skupa. Umesto toga, model mora naučiti robusnije feature-e koji generalizuju na varijacije.

- **Simulacija realnih varijacija**
  U produkciji, model će često morati da koristi zvukove koji se razlikuju od trening podataka po kvalitetu snimanja, pozadinskom šumu, akustici prostorije, karakteristikama govornika, itd. Augmentacijom možemo da simuliramo ove varijacije tokom treninga, čineći model otpornijim na realne uslove korišćenja.

- **Balansiranje skupa podataka**
  U mnogim scenarijima, dostupni podaci nisu ravnomerno distribuirani kroz klase. Na primer, u skupu podataka za klasifikaciju emocija govornika, normalan govor može biti mnogo češći od ljutitog ili tužnog govora. Selektivnom augmentacijom manje zastupljenih klasa možemo poboljšati balans skupa podataka.

4. Augmentacija audio signala u vremenskom domenu

4.1 Time Stretching

Time stretching je metoda augmentacije audio podataka kojom se produžava ili skraćuje trajanje audio signala bez promene visine tona.

Među jednostavnijim metodama za time stretching je SOLA - *Synchronous Overlap-Add*. Algoritam funkcioniše tako što deli audio na preklapajuće segmente, *frame*-ove i rekonstruiše audio signal sa drugačijim rastojanjem između segemenata. Kod usporenja to znači približavanje segmenata, dok kod produženja se segmenti udaljavaju. Radi na sledećem principu

1. **Deljenje na segmente**
   Signal se deli na preklapajuće prozore fiksne veličine, oko 10-100ms, zavisno od toga koliko želimo da promenimo dužinu.
2. **Preklapanje segmenata**
   Svaki segment se parcijalno prekriva drugim segmentom (izbor segmenta zavisi od sličnosti između njih, tj biramo slične segmente za preklapanje). Preklapanje se vrši kako bi ublažili prelaz između segmenata

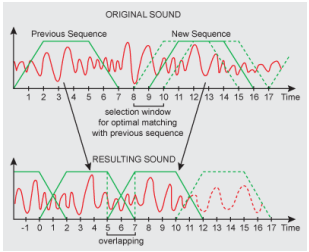


Figure 6: SOLA

Ovaj algoritam ima više varijanti kao što su *WSOLA* i *PSOLA*. U praksi, najbolji je kod jednostavnih, periodičnih audio signala. Nekada daje loše rezultate, kada dođe do lošeg preklapanja segmenata.

Bolja metoda time stretchinga je **Phase vocoder** algoritma koji radi na sledeći način:

1. **Podeli spectrogram na kratke vremenske prozore tj, *frame*-ove** Dobijeni *frame*-ovi se preklapaju delom.
2. **Primeni STFT na *frame*-ove**
   čime dobijamo amplitude i faze za sve frekvencije u svakom od preklapajućih *frame*-ova
3. **Promena rastojanja *frame*-ove**
   Kod ubrzavana *frame*-ovi se približavaju kreirajući overlap. Kod usporavanja će rastojanje biti veće, pa će nastati prazan prostor koji će se popuniti dupliranim *frame*-ovima ili interpolacijom.
4. **Prilagodi fazu svake frekvencije kako bi obezbedio kontinuitet**
   Kako bi izbegli primetne skokove u zvuku, potrebno je prilagoditi faze frekvencija kako bi se bolje usaglasili *frame*-ovi.
5. **Primeni inverzni STFT**
   Dobijamo izvorni zvuk ali usporen ili ubrzan.

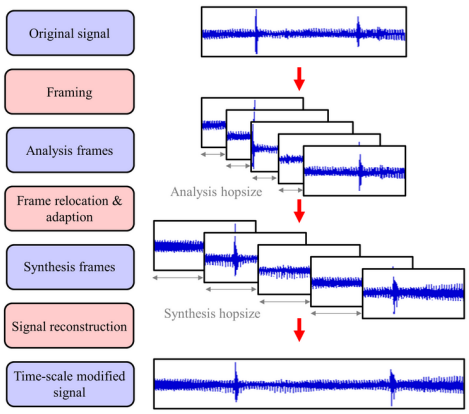


Figure 7: Phase vocoder

Koliko će se promeniti trajanje signala kontroliše se parametrom α.

- $\alpha > 1$ - ubrzan snimak
- $\alpha < 1$ - usporen
- $\alpha = 1$ - bez promene

Time stretching je koristan za

- Simulaciju različitih brzina govora
- Augmentaciju podataka za prepoznavanje govora
- Povećanje invarijantnosti na tempo kod muzike
- Prilagođavanje dužine audio snimaka

Prednosti	Nedostaci
Zadržava visinu i tonalitet	Povećava vremensku raznovrsnost podataka
Prirodno zvuči u razumnom opsegu faktora	Računski zahtevnije nego jednostavnije augmentacije
Povećava vremensku raznovrsnost podataka	Može narušiti prirodnost govora ako se pretera

Table 4: Prednosti i nedostaci time shiftinga

4.2 Pitch Shifting

Transformacija komplementarna time stretching-u. Njen cilj je da promeni visinu, tj pitch tona audio signala bez promene njegovog trajanja.

Za pitch shifting možemo koristiti granularnu sintezu ili već pomenuti phase vocoder.

Kod granularne sinteze

1. **Delimo signal na sitne intervale**
   Dok se krećemo kroz audio signal delimo ga na vrlo kratke *grain*-ove, dužine 1-100ms. Grain-ovi se obično preklapaju (overlap) da bi signal bio kontinualan.

2. **Menjamo gustinu reprodukcije *grain*-ova**
   Umesto menjanja trajanja samih grain-ova, menjamo **koliko često** ih reprodukujemo. Za viši ton, grain-ove reprodukujemo češće nego što su izvađeni iz originalnog signala. Za niži ton, reprodukujemo ih ređe.

3. **Korišćenje amplitude envelope**
   Svaki grain dobija amplitude envelope (najčešće Gaussian ili Hanning prozor) koji je nula na krajevima i maksimalan u sredini. Ovo omogućava glatko preklapanje grain-ova bez zvučnih artefakata.

4. **Overlap-add sinteza**
   Grain-ovi se sabiru (overlap-add) na izlazu, formirajući kontinualan signal. Veći stepen preklapanja daje glađi, prirodniji zvuk, dok manji overlap može dati "granularnu" teksturu.

5. **Održavanje trajanja (opciono)**
   Ako želimo da zadržimo originalnu dužinu snimka dok menjamo pitch:

   - Za pitch up: neki grain-ovi se dupliraju ili reprodukuju više puta
   - Za pitch down: neki grain-ovi se preskakuju/izbacuju

Druga opcija je koristiti phase vocoder kao i kod time stretchinga.

U ovom slučaju

- Primenimo time stretching za faktor $ $
- Resamplujemo signal sa faktorom $\frac{1}{\alpha}$
- Rezultat ima originalno trajanje ali promenjen pitch

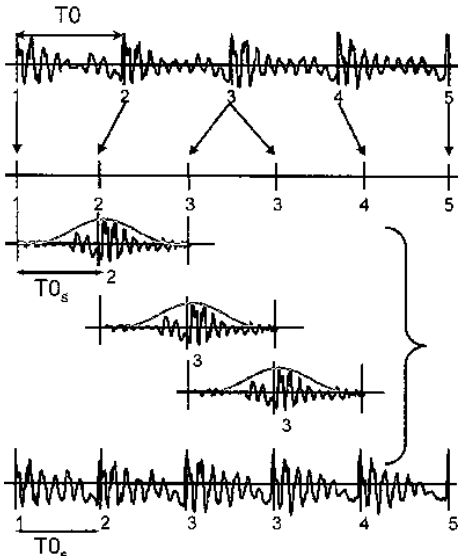


Figure 8: Pitch shifting

Primene pitch shifting-a su sledeće

- Simulacija različitih govornika (muški/ženski glasovi imaju različite fundamentalne frekvencije)
- Augmentacija za muzičku klasifikaciju (transpozicija)
- Kreiranje sintetičkih glasova
- Korekcija tona kod pevačkih aplikacija

Prednosti	Nedostaci
Uvodi varijabilnost u tonalitetu bez menjanja sadržaja	Artifakti pri većim shift-ovima
Pomoć pri simuliranju različitih govornika	Može promeniti percepciju pola govornika
Relativno prirodan zvuk u umerenom opsegu	

Table 5: Prednosti i nedostaci pitch shiftinga

4.3 Dodavanje nasumičnog šuma

Dodavanje šuma je jedna od najčešćih i najefektivnijih audio augmentacija. Simulira realne uslove snimanja i čini model robusnijim na različite vrste pozadinskog šuma.

Funkcioniše tako što se čist audio signal $x(t)$ kombinuje sa šumom $n(t)$

$$y(t) + x(t) + \alpha \dot{n}(t)$$

pri čemu je α intenzitet šuma. Često se koristi i SNR, tj *Signal to Noise Ratio*

$$SNR = 10 * log_{10}(P_{signal}/P_{noise})$$

gde su P_{signal} i P_{noise} prosečne snage signala i šuma.

Postoji više vrsta šuma, tj više distribucija šuma koje se mogu koristiti u augmentaciji.

- **Beli šum**
  - Šum koji ima podjednaku amplitudu među svim frekvencijama.
  - Zbog ovoga ljudskom uhu zvuči kao šuštanje.
  - Sample-ovi belog šuma su međusobno nezavisni i moguće je generisati ga sample-ovanjem bilo koje nezavisne distribucije (Gausian, Uniform …)
- **Roze šum**
  - Šum gde snaga signala opada sa porastom frekvencije.
  - Zato je šum dosta prirodniji za ljudsko uho i češće se javlja u prirodi.
  - Može da se dobije od belog šuma smanjivanjem snage sa porastom frekvencije
- **Brownian šum**
  - Šum gde snaga signala opada proporcionalno kvadratu frekvencije.
  - Još dublji u odnosu na roze šum.
- **Ambijentalni šum**
  - Šum nastao u realnim okruženjima
  - Najrealističniji ali zahteva eksterni skup podataka, teško generisati kao prethodne

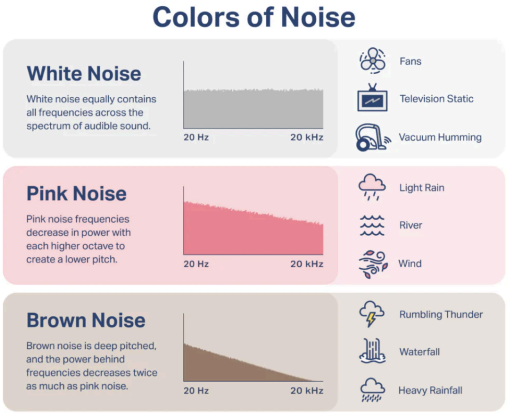


Figure 9: Vrste šuma

Prednosti	Nedostaci
Izuzetno efikasna za povećanje robusnosti	Previše šuma može narušiti razumljivost govora
Simulira realne uslove korišćenja	Potreban eksterni dataset kod ambijentalnog šuma
Jednostavna implementacija	Može maskirati važne karakteristike signala

Table 6: Prednosti i nedostaci dodavanja šuma

4.4 Time Shifting

Predstavlja jednostavnu transformaciju gde se audio signal pomera unapred ili unazad u vremenu. Na početku ili kraju se zato dodaje tišina ili se cirkularno rotira signal.

Primena time shiftinga - Simulacija različitih početnih tačaka detekcije - Augmentacija za sound event detection - Povećanje robusnosti na poziciju događaja u snimku

Prednosti	Nedostaci
Ekstremno brza transformacija	Može obrisati važan sadržaj na početku ili kraju
Ne menja akustički sadržaj	Manje korisna kada je temporalna struktura važna

Table 7: Prednosti i nedostaci time shiftinga

4.5 Random Gain

Random Gain je transformacija koja nasumično povećava amplitudu audio signala. Cilj ove transformacije je da simulira različite udaljenosti izvora zvuka od mikrofona ili različite postavke pojačanja tokom snimanja zvuka.

Matematički se definiše kao

$$y[n] = g \cdot x[n]$$

gde se g sempluje iz log-uniformne raspodele, s obzirom da ljudsko uho percipira glasnoću logaritamski.

Pritom je promena u decibelima data kao

$$\Delta dB = 20 \cdot \log_{10}(g)$$

Primena:

- Simulacija različitih udaljenosti od mikrofona
- Normalizacija varijacija u amplitudi između različitih snimaka
- Robusnost na različite postavke pojačanja

Prednosti	Nedostaci
Ekstremno jednostavna i brza	Previše pojačanja dovodi do clippinga
Prirodna varijacija koja postoji u realnim podacima	Bolja u kombinaciji sa drugim transformacijama

Table 8: Prednosti i nedostaci random gain-a

5. Augmentacija audio signala u frekventnom domenu

Augmentacije u frekvencijskom domenu rade na spektralnim reprezentacijama audio signala kao što su spectrogrami ili mel-spectrogrami. Ove tehnike često omogućavaju precizniju kontrolu nad specifičnim frekvencijskim karakteristikama.

5.1 Spectral masking (SpecAugment)

Spectral masking ili *SpecAugment* je napredna tehnika augmentacije audio podataka, prvi put opisana u radu Google Brain-a tima iz 2019. pod naslovom *SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition*. Ova metoda je pokazala dosta dobre rezultate kod automatskog prepoznavanje ljudskog govora (ASR).

Metoda je bazirana na augmentaciji log mel spektrograma ulaznog audio signala i inspirisana je maskiranjem kod augmentacije u vizuelnom domenu. Tehnika je komputaciono jeftina i primenjuje se direktno na log mel spektrogramu, kao da je spektrogram slika. Zbog ovoga je moguće primeniti spectral masking online, tokom treninga.

SpecAugment se sastoji od tri transformacije nad log mel spektrogramom

- Time warping
- Time masking
- Frequency masking

Time warping

Ukoliko imamo log mel spektrogram sa τ vremenskih koraka, mi ga posmatramo kao sliku gde je vreme horizontalna osa a frekvenca vertikalna osa. Nasumična tačka duž horizontalne linije koja prolazi kroz centar slike u okviru vremena $(W, \tau - W)$ će se *warp*-ovati ulevo ili udesno za distancu w koja se uzima iz uniformne raspodele u opsegu $[0, W]$. Deo signala levo od izabrane tačne se razvlači dok se desni deo skuplja.

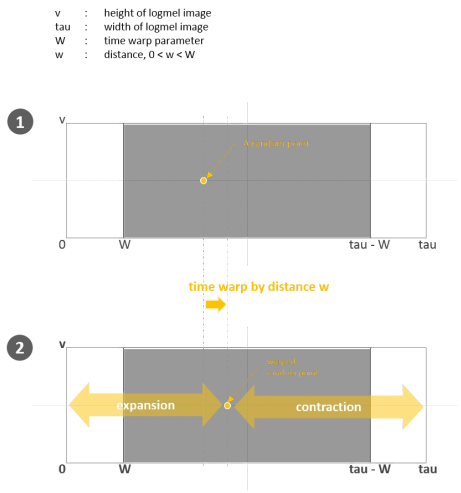


Figure 10: SpecAugment - Time warping

Frequency masking

Frequency masking se primenjuje tako da se f uzastopnih mel frequency kanala $[f_0, f_0+f]$ maskiraju, pri čemu se f bira iz uniformne rapsodele u opsegu $[0, F]$ (F je parametar) a f_0 se bira iz raspona $[0, \nu - f]$, gde je ν broj mel frequency kanala.

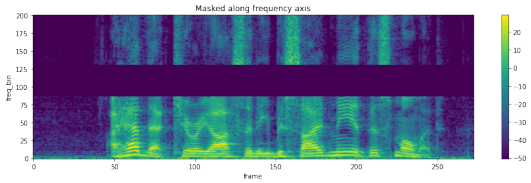


Figure 11: SpecAugment - Frequency masking

Time masking

Time masking se primenjuje tako da se t vremenskih koraka $[t_0, t_0 + t]$ maskiraju, pri čemu se t bira iz uniformne rapsodele u opsegu $[0, T]$ (T je parametar) a t_0 se bira iz raspona $[0, \tau - t]$, gde je τ broj vremenskih koraka signala.

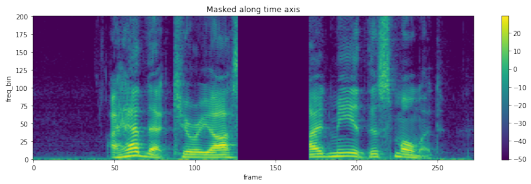
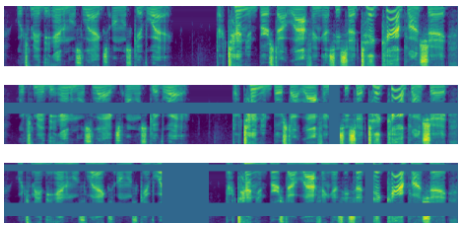


Figure 12: SpecAugment - Time masking

Moguće je definisati polise - kombinacije ove 3 metode, sa različitim vrednostima za W, F, i T, kao i primeniti ih na više raspona frekvenci ili više raspona vremenskih koraka. Neki od primera takvih polisa su iz orginalnog rada - LB, LD, SM i SS koji (izuzev LB) primenjuju transformacije nad 2 odvojena raspona frekvencija i 2 raspona vremenskih koraka, sa različitim W, F i T parametrima.



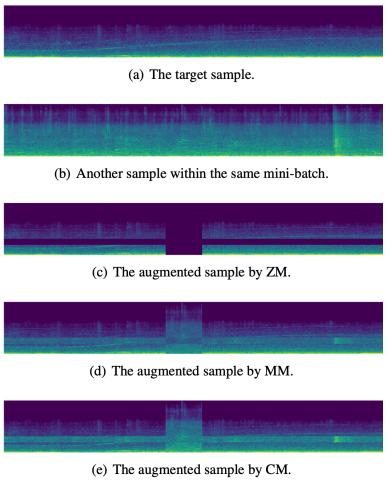
Policy	W	F	m_F	T	p	m_T
None	0	0	-	0	-	-
LB	80	27	1	100	1.0	1
LD	80	27	2	100	1.0	2
SM	40	15	2	70	0.2	2
SS	40	27	2	70	0.2	2

Figure 13: SpecAugment - Policies

Pored klasičnog SpecAugmenta, postoji i druge varijante koje su nastale iz ove metode. Prvi nedostatak kod SpecAugment-a je što se augmentacija podataka vrši samo nad ulaznim podacima neuronske mreže, što propušta priliku za augmentacijom mel spektrograma *hidden state*-a neuronske mreže. Upravo ovo je ideja algoritma *SpecAugment++*.

Pored toga, ova tehnika uvodi i mogućnost maskiranja frekvenci i vremena, ali ne striktno anuliranjem, već i dodeljivanjem druge vrednosti. Konkretno, rad je opisao 3 metoda augmentacije.

- **ZM - Zero Value Masking**
  Klasična transformacija, kao i kod *SpecAugment*a. Deo frekvencija i deo vremenskih koraka se postavlja na 0.
- **MM - Mini-batch based mixture masking**
  Kod ove metode se uzimaju mini-batch-evi audio signala nad kojima se primenjuje augmentacija. Tada se hidden state trenutnog *sample*-a $x \in \mathbb{R}^{T \times X}$ menja tako što se uzima hidden state nekog drugog *sample*-a iz istog mini-batcha $y \in \mathbb{R}^{T \times X}$ i u okviru odabranog raspona frekvenci i odabranog raspona vremenskog intervala se računa vrednost mel spektrograma uzorka x kao prosek vrednosti x i y.
- **MM - Mini-batch based mixture masking**
  Kod ove metode se uzimaju mini-batch-evi audio signala nad kojima se primenjuje augmentacija. Tada se hidden state trenutnog *sample*-a $x \in \mathbb{R}^{T \times X}$ menja tako što se uzima hidden state nekog drugog *sample*-a iz istog mini-batcha $y \in \mathbb{R}^{T \times X}$ i u okviru odabranog raspona frekvenci i odabranog raspona vremenskog intervala mel spektrogramu uzorka x se dodeljuje vrednost spektrograma y.



(a) The target sample.

(b) Another sample within the same mini-batch.

(c) The augmented sample by ZM.

(d) The augmented sample by MM.

(e) The augmented sample by CM.

Figure 14: SpecAugment++

Najveći nedostatak ove metode je potreba za pažljivi tune-ovanjem parametara augmentacije. Zbog toga je nastala varijacija ovih metoda koja pokušava da automatski odredi najbolje parametre - *Policy-SpecAugment*.

Neka su moguće augmentacije $\{A_1, A_2, A_3\} = \{TimeWarp, FreqMask, TimeMask\}$ i neka je u epohi j, prosečan validation loss primenom samo i-te augmentacije $\mathcal{L}_i^j$.

Zatim se ti validation loss-ovi koriste kako bi se dobio vektor verovatnoća

$$P_i^{(j)} = \frac{\mathcal{L}_i^{(j)}}{\sum_{k=1}^{3} \mathcal{L}_k^{(j)}} \tag{3}$$

Ovako dobijenim vektorima verovatnoće moguće je dalje usmeriti augmentaciju tako da se primenjuju one transformacije koje će poboljšati performanse.

Prednosti	Nedostaci
Izuzetno efikasna za ASR zadatke	Može maskirati kritične informacije
Može maskirati kritične informacije	Potrebno pažljivo tune-ovanje parametara za specifične zadatke
Brza jer radi direktno na spektrogramu	

Table 9: Prednosti i nedostaci SpecAugment-a

5.2 Loudness kontrola

5.3 Filtering

6. Napredne tehnike augmentacije

7. Evaluacija augmentacije