UO‘K 004.934.8

NUTQ SIGNALLARI ASOSIDA SO‘ZLOVCHILARNI TANIB OLISH ALGORITMLARINING SAMARADORLIGINI BAHOLASH

[+] *Shukurov K.E.*[1], *Xasanov U.K.*[1]

[1] Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti, Toshkent, O‘zbekiston

[+] keshukurov@gmail.com

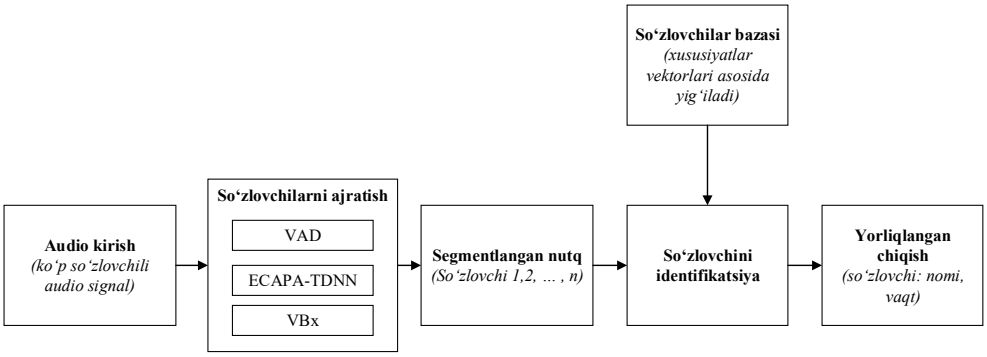
Annotatsiya. Ushbu maqolada so‘zlovchilarni tanib olsih jarayonlarida turli modellardan foydalanishning samaradorliklarini baholash va tizim uchun eng yaxshisini tanlash tahlil qilingan. Tizimning aniqlik va tezlik samaradorliklarida jihatidan tahlil qilishda klassik MFCC + kosinus o‘xshashlik va zamonaviy x-vector, ECAPA-TDNN + PLDA arxitekturalar o‘zaro taqqoslanadi. Turli so‘zlovchilar asosida hosil qilingan ma’lumotlar to‘plami asosida modellarning aniqlik, f1-score, EER, kechikish va GPU yuklamalari ko‘rsatkichlari baholangan. Tajriba natijalari bo‘yicha ECAPA-TDNN modeli 95.7 % aniqlik bilan qolgan modellardan ustunligini ko‘rsatdi. So‘zlovchilarni ajratish tizimlari uchun ham so‘zlovchilarni aniqlash bosqichi muhim ahamiyat kasb etganligi uchun aniqlik ko‘rsatkichlari yuqori darajadagi dolzarblikka ega. Hisoblash resurlaridan foydalanish, katta ma’lumotlar to‘plamlari bilan ishlash hamda ularning ma’lumotlarni tahlil qilishda ECAPA-TDNN + PLDA modeli yaxshi yechimlarni taklif etadi.

Kalit so‘zlar: so‘zlovchilarni identifikatsiya qilish, ECAPA-TDNN, x-vector, MFCC, log-Mel, PLDA, chuqur o‘qitish, xususiyatlar vektori, nutq biometrikasi, AM-softmax.

1 KIRISH

Globallashuv jarayonida axborot texnologiyalari bilan bir qatorda inson-mashina interfeyslari ham tez rivojlanmoqda. Inson mashina interfeyslarining sun’iy intellekt algoritmlari bilan integratsiyasi yuqori samaradorlik kasb etadi [1]. Bunday tizimlarni ishlab chiqishda nutq eng qulay vositalardan biri hisoblanadi. Nutq yordamida buyruqli muloqotni shakllantirishda nutqni matnga aylantirish, so‘zlovchilarni tanib olish texnologiyalaridan foydalaniladi [2-5]. Nutqni matnga aylantirish tizimlari bilan bir qatorda so‘zlovchilarni aniqlash inson-mashina muloqot jarayonlarining muhim tarkibi qismi sifatida qaraladi. So‘zlovchilarni tanib olish insonning nutq signallari orqali so‘zlovchini kim ekanligini aniqlash jarayoni bo‘lib, biometrik identifikatsiya, xavfsizlik tizimlari, aloqa-markazlari, sud-ekspertizasi, intellektual yordamchilar kabi ko‘plab sohalarda keng qo‘llaniladi [6].

Ko‘p so‘zlovchili nutq signallariga intellektual ishlov berish tizimlari so‘zlovchilarning nutqni avtomatik aniqlab olish dolzarb muammolardan biri. So‘zlovchilarni ajratish va identifikatsiya qilish asosidagi integratsiyalashgan tizim bu muammolarni yechishga yordam beradi. 1-rasmda bu ikki modulning integratsiyalashgan holda ko‘p so‘zlovchili nutqdan identifikatsiya qilingan so‘zlovchilarni ketma-ketlik nutqlarini aniqlash bosqichlari keltirilgan [7-9].



1-rasm. So‘zlovchilarni ajratish va identifikatsiyalash tizimining umumiy arxitekturasi

Nutq signallari orqali so‘zlovchilarni tanib olishda dastlab GMM-UMB yondashuvlari asosida amalga oshirilgan, keyinchalik i-vector yondashuvlari asosida tanib olish modellari rivojlandi [10]. Chuqur o‘qitish algoritmlarining rivojlanishi natijasida x-vector, ECAPA-TDNN, Res-Net kabi yondashuvlar assida modellar paydo bo‘ldi [10-12]. Bu modellar katta hajmdagi ochiq nutq korpuslarini ishlab chiqishda, ularni sinovdan o‘tkazishda va samaradorlikni baholashda foydalaniladi [13].

Zamonaviy so‘zlovchilarni tanib olish tizimlari so‘zlovchini ifodalovchi xususiyatlar vektorlari asosida ishlaydi. Nutq signalidan ajratib olingan xususiyatlar neyron tarmoqlaridan olingan individual xususiyatlar vektorlari shaklida ifodalanadi, hamda ular orasida o‘xshashlik mezonlari asosida solishtiriladi. So‘zlovchini tanib olish tizimlari ko‘p so‘zlovchili, real vaqt tizimlar, ovoz interfeyslarini takomillashtirish jarayonlarida samarali natijalarni beradi [14].

2 ASOSIY QISM

So‘zlovchilarni tanib olish nutq signallariga dastlabki ishlov berish boshqichidan boshlab uni aniqlash darajasigacha olib chiqadigan statistik va akustik bosqichlarni o‘zida mujassamlashtiradi. Ushbu tizim dastlabki ishlov berish, akustik xususiyatlarni ajratish, chuqur neyron tarmoq asosida so‘zlovchini ifodalovchi xususiyatlar vektorlarini hosil qilish hamda olingan xususiyatlar vektorlar asosida o‘xshashlikni baholash bosqichlaridan tarkib topgan [13]. 2-rasmda keltirilgan chizmadagi har bir bosqichning aniqlik, tezkorlik, barqarorlik kabi samaradorlikni baholash mezonlari umumiy tizimning samaradorligiga ta’sir qiladi.

2-rasm. So‘zlovchini aniqlash tizimining umumiy arxitekturasi

Dastlabki ishlov berish bosqichida signal sifati yaxshilash, turli xalaqitlar va shovqinlarni kamaytirish bajariladi [16]. Bu bosqichning muhimligi signalning akustik tahlil jarayonlari uchun barqaror holatga keltiriladi. Xususiyatlarni ajratish bosqichida nutq chastota sohasiga o‘tkaziladi, inson eshitish modeliga yaqin akustik ko‘rsatkichlari olinadi. Bu bosqichda MFCC, log-Mel ko‘rsatkichlari, PLP koeffitsiyentlari aniqlanadi [17]. So‘zlovchilarni ifodalovchi xususiyatlar vektori hosil qilish bosqichida chuqur neyron tarqmoq modellaridan foydalaniladi [18]. Baholash bosqchida tizim yangi nutq namunasi xususiyatlar vektorlarini avvaldan hosil qilingan so‘zlovchini ifodalovchi xususiyatlar vektorlari asosidagi ma’lumotlar to‘plami bilan taqqoslaydi.

2.1 Xususiyatlarni ajratish

So‘zlovchilarni tanib olish tizimlarining eng muhim bosqichlaridan biri bu xususiyatlarni ajratish jarayoni hisoblanadi. Bu bosqichda xususiyatlarni aniq ajratish tizimning samaradorligiga sezilarli ta’sir qiladi [21]. Nutq signali inson tovush hosil qilish tizimlari orqali hosil qilinadi. Shuning uchun u eshitish tizimining chastotaviy sezgirligi bilan uzviy bog‘liq. Olib borilgan tadqiqotlar shuni ko‘rsatadiki inson qulog‘i juda sezgir bo‘lgan chastotalar oralig‘ini (300-3400 Hz) qayta taqdim etuvchi Bark yoki Mel shkalalari asosidagi filtrlar qo‘llaniladi [22]. Buning natijasida MFCC, log-Mel yoki PLP kabi belgilar akustik ma’lumotlarni juda ixcham formatda ifodalaydi.

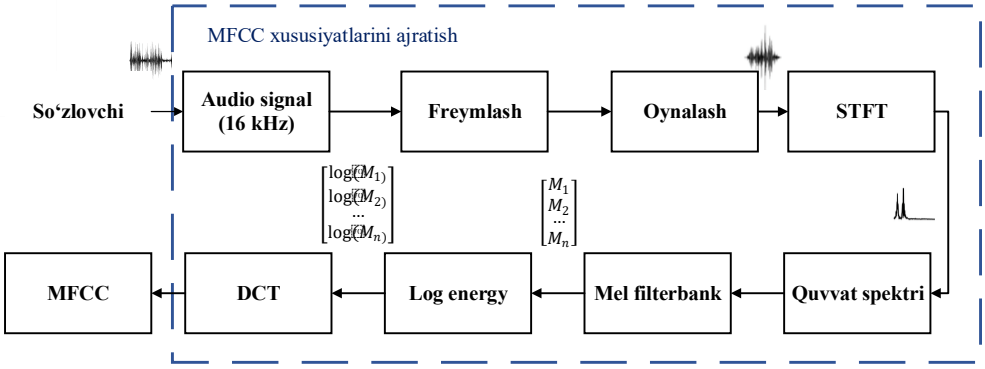
MFCC koeffitsiyentlari nutq signallarining chastota tarkibiy qismlarini logarifmik o‘lchovda tasvirlaydi hamda diskret kosinus o‘zgartirish (DCT) yordamida yuqori o‘lchamdan juda ixcham ko‘rinishga o‘tadi. 3-rasmda keltirilgan MFCC koeffitsiyentlarini olish bosqichlari keltirilgan [23].

STFT (Short-Time Fourier Transform) algoritmi orqali nutq signal spektral domeniga o‘tkaziladi:

$$X_m(\omega)=\sum_{n=0}^{L-1}x[n+mH]\cdot w[n]\cdot e^{-j\omega n}, \tag{1}$$

bu yerda L – kadr uzunligi, H – qadam kattaligi, $w[n]$ – Hamming oynasi, quvvat spektri hisoblanadi:

$$P_m(f)=\frac{|X_m(f)|^2}{L}. \tag{2}$$



3-rasm. So‘zlovchini aniqlash tizimining umumiy arxitekturasi

Mel-filtrlar banki yordamida chastotalar Mel shkalasiga o‘tkaziladi:

$$E_{m,b} = \sum_f P_m(f) \cdot H_b(f). \tag{3}$$

Logarifmik amplituda olinadi hamda so‘nggi bosqich uchun DCT qo‘llaniladi, natijada Mel chastotali kepstral koeffitsiyentlarini olinadi:

$$F_{m,b} = \log\left(E_{m,b} + \varepsilon\right), \tag{4}$$

$$MFCC_{m,c} = \sum_{b=1}^{B} F_{m,b} \cos\frac{\pi c\left(b-0.5\right)}{B}, \quad c = 0,1,\ldots,C-1. \tag{5}$$

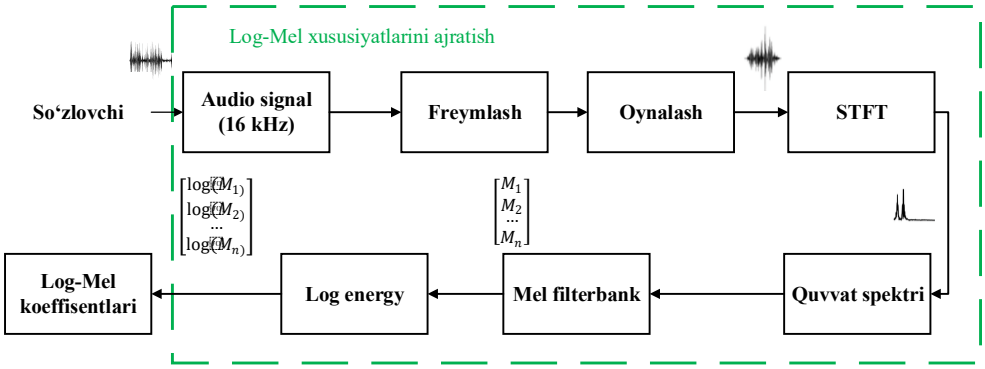
Nutq signallariga intellektual ishlov berish jarayonlarida 20-25 millisekund kadrlar olinadi, har bir kadr asosida 13-23 tagacha MFCC koeffitsiyentlari olinadi. Ayrim hollarda aniqlik darajalarini oshirish uchun MFCC koeffitsiyentlarining birinchi hamda ikkinchi tartibli hosilalari olinadi, chunki bu hosilalar signalning vaqt dinamikasini ifodalaydi.

2.2 Log-Mel filtrbank koeffitsiyentlari

Chuqur neyron tarmoqlarning rivojlanishi bilan MFCC koeffitsiyentlari o‘rnigan log-Mel koeffitsiyentlaridan foydalanish ommalashdi, bu koeffitsiyentlar xuddi MFCC koeffitsiyentini olish bosqichlari bilan bir xil, faqat DCT bosqichini o‘z ichiga olmaydi [24]:

$$F_{m,b} = \log\left(\sum_f P_m(f) \cdot H_b(f)\right). \tag{6}$$

Log-Mel energiyalari xom spektr shakliga yaqin bo‘lib, ular logarifmik o‘lchovda bo‘lgani uchun inson eshitish tizimining sezgirligiga mos hamda x-vector, ECAPA-TDNN kabi modellar uchun ma’lumotlarning yaxlit spektral tuzilmasini saqlaydi.



4-rasm. Log-Mel koeffitsiyentlarini olish jarayonining ketma-ketligi

2.3 Chuqur o‘qitish asosidagi modellar

So‘zlovchilarni aniqlash tizimlarini ishlab chiqishda chuqur o‘qitish modellarini joriy etish tizim samaradorligini oshishiga olib keldi. Klassik modellardan farqli ravishda chuqur neyron tarmoqlari asosidagi modellar so‘zlovchilarni aniqlashda ularni farqlashga qaratilgan yuqori darajadagi so‘zlovchini ifodalovchi xususiyatlar vektorlarini o‘qitiladi [25].

X-vector modeli TDNN arxitekturasiga asoslangan bo‘lib, ushbu tarmoq har bir nutq segmentlarining vaqtli kontekstlarini hisobga olgan holda, har kadrga tegishli akustik belgilarning ketma-ketligidan segmentlar darajasida xususiyatlar vektorini hosil qiladi. Uning asosiy arxitekturasi bir nechta bosqichdan iborat [25]:

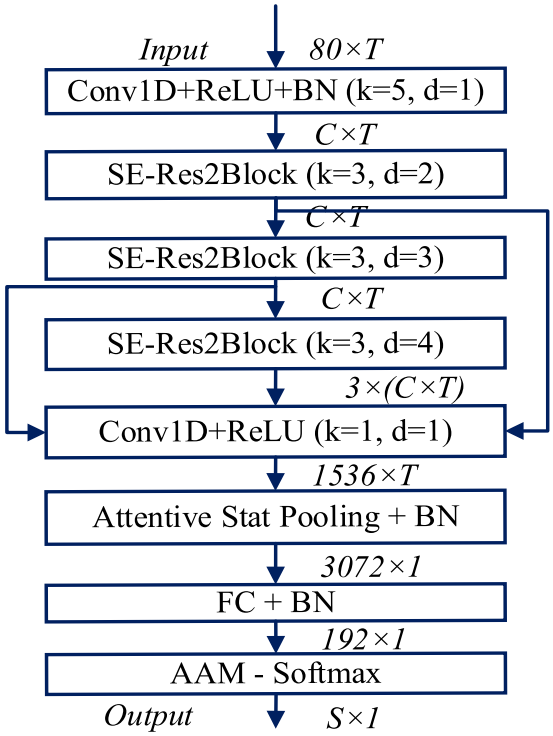
1. Nutq signallari uzluksiz bo‘lgani sababli ishlov berish jarayonlarida kadrlarga bo‘linadi. TDNN qatlamlari har bir kadrdagi ma’lumotlardan tashqari avvalgi va keyingi kadrlardagi ma’lumotlarni ham tahlil qiladi;
2. Statistik birlashtirish bosqichida model barcha kadrlar natijalarini bir joyga mujassamlashtiradi:

$$\mu = \frac{1}{T}\sum_{t=1}^{T} h_t,\ \sigma = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(h_t - \mu)^2}, \tag{7}$$

bu yerda μ – o‘rtacha qiymat, σ – dispersiya;

3. Avvalgi bosqichdan olingan ma’lumotlar tarmoqning keyingi qatlamlariga uzatiladi, bu qatlamlar ma’lumotlarni qayta ishlaydi hamda har bir so‘zlovchi uchun alohida xususiyat vektorlarini (x-vector) hosil qiladi.

ECAPA-TDNN modeli x-vector arxitekturasining takomillashtirilgan varianti hisoblanib, modelning asosiy farqlaridan biri u kanallararo diqqat mexanizmi (ing: Squeeze and Excitation), Res2Net bloklari va e’tibor mexanizmi asosidagi statistik yig‘ish (ing: attentive statistics pooling) qatlamlarini birlashtirgan murakkab arxitektura tuzilishiga ega [26]. 5-rasmda x-vectorning takomillashtirilgan ECAPA-TDNN yondashuvning asosiy bosqichlari keltirilgan.



5-rasm. ECAPA – TDNN modelining arxitekturasi

Model chiqishida hosil bo‘ladigan so‘zlovchini ifodalovchi xususiyatlar vektorlari 192-512 ta o‘lchamli bo‘lib, ular AM-Softmax yo‘qotish funksiyasi yordamida o‘qitiladi:

$$L = -\frac{1}{N}\sum_{i}\log\frac{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}}{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)} + \sum_{j\neq y_i} e^{s\cos\left(\theta_j\right)}}, \tag{8}$$

bu yerda m – margin (0.2-0.3), s – miqyos omili, θ_{y_i} – haqiqiy so‘zlovchi uchun burchak. Ushbu yo‘qotish funksiyasi xususiyatlar vektorlarini sferik sohada ajratib, so‘zlovchilar orasidagi farqlarni maksimal, o‘xshashliklarni esa minimal qiladi.

2.4 So‘zlovchilarni aniqlash algoritmlari

So‘zlovchilarni aniqlash algoritmlari umumiy tizimning yakuniy bosqichi bo‘lib, ushbu bosqichda so‘zlovchini ifodalovchi xususiyatlar vektorlari o‘zaro taqqoslanadi. Bu jarayonni amalga oshirishda kosinus o‘xshashlik, PLDA va boshqa algoritmlardan foydalaniladi [14].

Kosinus o‘xshashlik eng oddiy va tezkor usul bo‘lib, ikki xususiyatlar vektorlari orasidagi burchak masofasi o‘xshashlikni aniqlaydi.

$$Score_{cos}\left(E_1, E_2\right) = \frac{E_1 \cdot E_2}{E_1 E_2}. \tag{9}$$

PLDA esa ehtimolliklar asosidagi sinflashtirish modeli bo‘lib, xususiyatlar vektorlari fazosini so‘zlovchilararo dispersiyalarga ajratadi:

$$E = \mu + V \cdot y + \varepsilon, \tag{10}$$

bu yerda E – so‘zlovchini ifodalovchi xususiyatlar vektori, μ – umumiy o‘rtacha vektori, V – latent (yashirin) so‘zlovchi faktorlari matritsasi, y – so‘zlovchiga xos yashirin vector, ε – shovqin tarkibiy qismlari.

Bayesian klassifikatori Bayes ehtimolliklar nazariyasiga asoslanadi va asosiy maqsad berilgan so‘zlovchini ifodalovchi xususiyatlar vektorlari uchun qaysi so‘zlovchi eng katta ehtimollikka ega ekanligini aniqlashdir [27].

$$P\left(s_i|E\right) = \frac{P\left(E|s_i\right) \cdot P\left(s_i\right)}{P\left(E\right)}, \tag{11}$$

bu yerda $P(E|s_i)$ – berilgan so‘zlovchiga tegishli xususiyatlar vektorlar ehtimoli, $P(s_i)$ – so‘zlovchining oldindan ehtimoli, P(E) – barcha sinflar bo‘yicha normalashuvchi omil.

3 BAHOLASH MEZONLARI VA MA’LUMOTLAR TO‘PLAMI

So‘zlovchini aniqlash tizimining samaradorligini hisoblashda turli xil baholash mezonlari qo‘llaniladi. Baholash jarayonlari odatiy tizimlar uchun aniqlik bilan bir qatorda boshqa baholash mezonlari o‘rtasidagi muvozanat ham tahlil qilinadi, chunki so‘zlovchilarni aniqlash jarayoni biometrik autentifikatsiya muhitida ishonchlilik va xavfsizlik talablariga bir vaqtning o‘zida javob berishi kerak. So‘zlovchini aniqlash tizimlarining eng muhim baholash mezonlaridan biri tenglik xato darajasi (ing: Equal Error Rate - EER) hisoblanadi. EER qiymati bu noto‘g‘ri tan olish ehtimoli (ing: false acceptance rate - FAR) va noto‘g‘ri rad etish ehtimoli (ing: false rejection rate - FRR) qiymatlari o‘zaro teng bo‘lgan nuqta, bu mezon tizimning umumiy xatolik darajasini belgilaydi, uning qiymati qancha past bo‘lsa model shuncha aniq ishlaydi [28]:

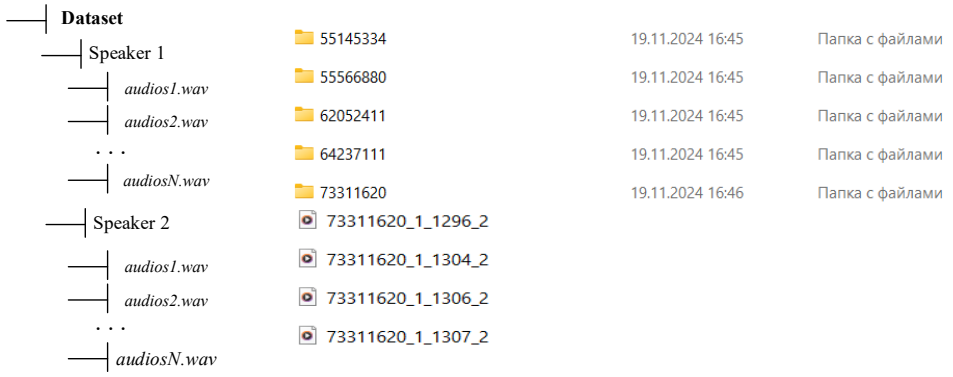
$$EER = \frac{FAR + FRR}{2}. \tag{12}$$

Tizim samaradorligini baholashda minimal aniqlashning xarajati funksiyasi (ing: minimum Detection Cost Function - minDCF) mezonidan ham foydalaniladi. Bu ko‘rsatkich NIST Speaker Recognition Evalution (SRE) protokollari asosida qabul qilingan standart bo‘lib, u xatolik narxi (cost) va foydalanish ehtimolliklarini hisobga olgan holda umumiy aniqlash qiymatini beradi [28].

$$C_{det} = C_{FA} \cdot P_{FA} \cdot P_{targ} + C_{miss} \cdot P_{miss}\left(1 - P_{targ}\right), \tag{13}$$

bu yerda C_{FA} – noto‘g‘ri qabul qilish narxi, C_{miss} – noto‘g‘ri rad etish narxi, P_{FA}, P_{miss} – mos ehtimollar, P_{targ} – haqiqiy so‘zlovchi paydo bo‘lish ehtimoli.

So‘zlovchilarni aniqlash tizimi uchun nutq signallari hamda ularning so‘zlovchilarni avvaldan belgilangan annotatsiyali ma’lumotlar to‘plami shaklantiriladi, har bir audio signalning diskretlash chastotasi 16kHz va uning uzunligi turlicha o‘lchamda bo‘ladi, bundan tashqari audio signallarning so‘zlovchilarining soni ham 100 dan oshiq bo‘lishi maqsadga muvofiq. Audio signallarning tarkibida turli muhitlarda yozib olingan nutq signallari, ovozlarning ham turli o‘lchamlarda bo‘lishi tanib olishda olib boriladigan ishni osonlashtiradi.



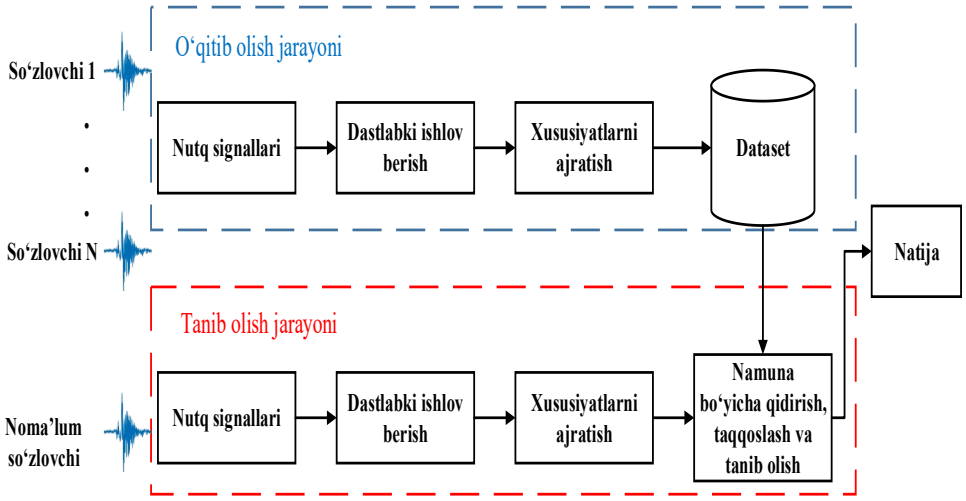
6-rasm. Ma’lumotlar to‘plamining umumiy strukturasi

6-rasmda keltirilgan dataset tuzilmasida har bir so‘zlovchi uchun alohida kataloglar ajratilgan bo‘lib, ularning ichida shu so‘zlovchiga tegishli bo‘lgan bir nechta .wav formatidagi audio fayllar joylashgan. Model o‘qitish jarayonlarida katalog nomlari so‘zlovchining nomi sifatida olinadi. So‘zlovchini aniqlash jarayonlarida yangi ovoz namunasi berilganda tizim uni qaysi katalogda ovozga mos kelishini aniqlaydi.

Har bir katalog ichida bir so‘zlovchiga tegishli bir nechta audio fayllar bo‘lganligi sababli, so‘zlovchini tekshirish jarayonlarida ikki audio yozuvning bir shaxsga tegishli yoki tegishli emasligi amalga oshiriladi. Bundan tashqari so‘zlovchilarning alohida kataloglarda saqlangan individual yozuvlari so‘zlovchilarni ajratish tizimlarini tayyorlashda namuna bazasi sifatidan ham foydalaniladi.

4 TAJRIBA SINOV NATIJALARI

7-rasmda so‘zlovchini identifikatsiya qilish tizimining umumiy strukturasi keltirilgan. So‘zlovchini tanib olish tizimining samaradorligini baholashdan asosiy maqsad ishlab chiqilgan modelning qay darajada aniqligi, tezligi hamda hisoblash resurslardan foydalanish samaradorligi ko‘rsatkichlarini obyektiv baholashdan iborat.



7-rasm. So‘zlovchini identifikatsiya qilish tizimi

Nutq signallarini qayta ishlash jarayonlarida shovqinlarni kamaytirish uchun gibrid filtrlar va normalizatsiya algoritmlaridan foydalaniladi. So‘ngra asosiy uch xil arxitektura uchun turli xususiyatlar ajratilib olinadi:

1. MFCC +cosine similarity uchun MFCC koeffitsiyentlari hamda ularning statistik xususiyatlari olinadi;
2. X-vector modeli uchun pretrained modeli asosida TDNN arxitekturasi orqali xususiyatlarni vektorlarini shakllantiriladi;
3. ECAPA-TDNN modeli uchun chuqur konvolutsion qatlamlar hamda SE-bloklardan foydalangan holda so‘zlovchini ifodalovchi xususiyalar vektori shakllantiriladi.

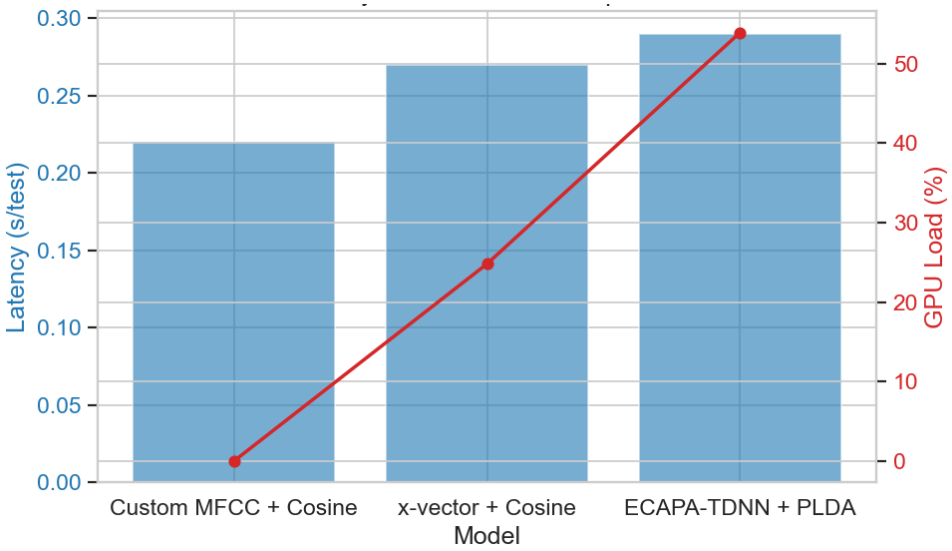
Har bir model uchun sinov audio fayllarining xususiyatlar vektorlari avvalda hosil qilingan dataset bilan solishtiriladi. So‘zlovchini aniqlash qarorlarini esa cosine similarity yoki PLDA orqali amalga oshiriladi. Baholash mezonlari sifatida esa accuracy, f1-score, EER va boshqalar olinadi.

1-jadval. So‘zlovchini identifikatsiya qilish modelining baholash ko‘rsatkichlari

Model	Aniqlik	F1-score	EER(%)	Latency(s/test)	GPU load (%)
MFCC+cosine similarity	91.2	0.937	9.8	0.22	-
X-vector+cosine similarity	93.8	0.952	5.6	0.29	24.8
ECAPA-TDNN+PLDA	95.7	0.971	3.9	0.27	54

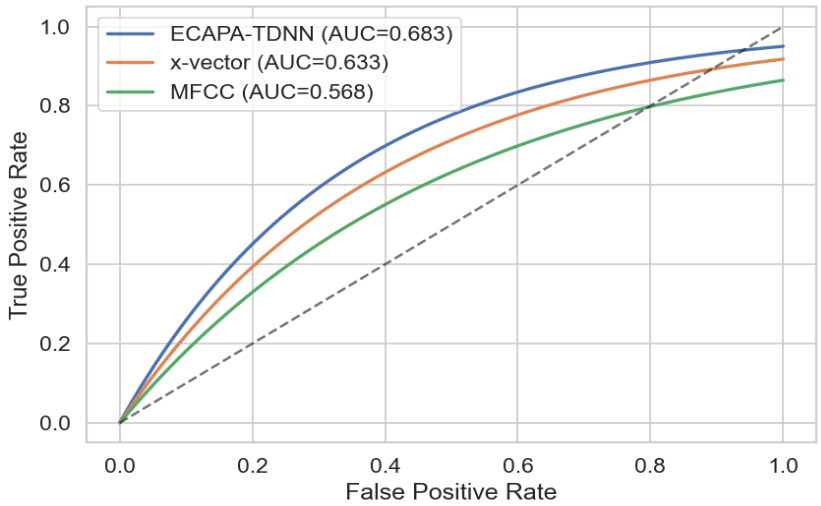
1-jadvalda turli so‘zlovchini aniqlash modellari asosida o‘tkazilgan tajriba natijalari keltirilgan bo‘lib, har bir model 100 dan ortiq so‘zlovchidan iborat bo‘lgan o‘zbek tilidagi maxsus datasetdan foydalanilgan. Jadvalda aniqlik (Accuracy), F1-score, xato aniqlash darajasi (EER), hisoblash kechikishi (Latency) va GPU yuklanishi (GPU Load) kabi asosiy samaradorlik ko‘rsatkichlari solishtirilgan.

ECAPA-TDNN hamda PLDA kompbinatsiyasi asosida ishlab chiqilgan model aniqlik ko‘rsatkichlari bo‘yicha yuqori natijalarni qayd etdi. Aniqlik 95.7% ga cha oshgan bundan tashqari modelning ishlab chiqish parallel hisoblash usullaridan foydalanilganligi uchun ham GPU uchun 54 % yuklama tushgan. So‘zlovchilarnin sonini oshishi model uchun qo‘shimcha yuklamalarga sabab bo‘ladi, bu esa o‘z navbatida chuqur o‘qitish modellari asosida ishlashni talab qiladi.



8-rasm. So‘zlovchini aniqlash arxitekturalarining hisoblash samaradorligini kechikish va GPU yuklamasini taqqoslash

8-rasmda uchta dinamikni identifikatsiyalash modellari uchun hisoblash kechikishi va GPU-dan foydalanish o‘rtasidagi muvozanat shakli ko‘rsatilgan. ECAPA-TDNN + PLDA modeli yuqori GPU yukiga (≈54%) va biroz yuqori hisoblash kechikishiga (0,29 s/sinov) ega. Biroq, yuqori aniqlik va katta ma’lumotlar hajmini talab qiladigan tizimlar uchun ECAPA-TDNN + PLDA yaxshiroq natijalar beradi.



9-rasm. So‘zlovchini aniqlash modellarining ROC egri chiziqlari (ECAPA-TDNN, x-vector, MFCC)

9-rasmda haqiqiy ijobiy va noto‘g‘ri ijobiy qiymatlar o‘rtasidagi bog‘liqlik ko‘rsatilgan. Egri chiziq ostidagi maydon modelning umumiy aniqligini ko‘rsatdi. Natija qancha yuqori bo‘lsa shuncha yaxshi. ECAPA-TDNN+PLDA AUC = 0,683 bilan eng yaxshi natijani ko‘rsatdi, x-vektor va MFCC + kosinus o'xshashligi mos ravishda AUC = 0,633 va AUC = 0,568 bilan natijalarni berdi.

5 XULOSA

Ushbu tadqiqot ishida so‘zlovchini aniqlash tizimlarining turli arxitekturalari tahlil qilinib ularnin samaradorliklari baholanishga qaratilgan yondashuv ishlab chiqildi. Nutq signallariga intellektual ishlov berish bosqichlariga jumladan xususiyatlarga ajratish, xususiyatlar vektorlarini hosil qilish va baholash bosqichlarida batafsil o‘rganildi. Tajribalar asosida ECAPA-TDNN + PLDA modeli 95.7 % aniqlik va 3.9 % EER bilan yuqori natijalarga erishdi. Bundan tashqari GPU resurslaridan 54% darajada samarali foydalanilib, 0.27 s/test kechikish bilan aniqlik hamda tezlik o‘rtasida muvozanatni taqdim etdi. So‘zlovchilarni ajratish hamda ularni aniqlash integratsion tizimlarini ishlab chiqishda aniqlik va tezlik o‘rtasida muvozanat yaxshi ta’minlanishi kerak. Bundan tashqari so‘zlovchilarni ajratish jarayonlarida ham ECAPA-TDNN asosidagi xususiyatlar vektorlari yaxshi natijalarni qayd etgani uchun integratsion tizimni amalga oshirishda shu modelni tavsiya etiladi.

ADABIYOTLAR

[1] *Azofeifa, Jose & Noguez, Julieta & Ruiz-Loza, Sergio & Molina Espinosa, José & Magana, Alejandra & Benes, Bedrich*. (2022). Systematic Review of Multimodal Human–Computer Interaction. Informatics. 9. 13. 10.3390/informatics9010013.

[2] *Shukurov K., Khasanov U., Turaev B., Kakhkharov A.* (2023). The Effectiveness of the Implementation of Speech Command Recognition Algorithms in Embedded Systems. The Eurasia Proceedings of Science Technology Engineering and Mathematics, 23, 220-224. https://doi.org/10.55549/epstem.1365795

[3] *S. Kamoliddin Elbobo ugli, K. Shokhrukhmirzo Imomali ugli and K. Umidjon Komiljon ugli.* "Uzbek speech commands recognition and implementation based on HMM," 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT), Tashkent, Uzbekistan, 2020, pp. 1-6, doi: 10.1109/AICT50176.2020.9368591.

[4] *Musaev M., Khujayarov I., Ochilov M.* (2023). Speech Recognition Technologies Based on Artificial Intelligence Algorithms. In: Zaynidinov, H., Singh, M., Tiwary, U.S., Singh, D. (eds) Intelligent Human Computer Interaction. IHCI 2022. Lecture Notes in Computer Science, vol 13741. Springer, Cham. https://doi.org/10.1007/978-3-031-27199-1_6

[5] *K. Shukurov, T. Boburkhon and U. Khasano.* "Implementation of speech processing algorithms based on Singular Value Decomposition and Hidden Markov Model," 2021 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2021, pp. 01-03, doi: 10.1109/ICISCT52966.2021.9670357.

[6] *Musaev M., Abdullaeva M., & Ochilov M.* (2022, August). Advanced feature extraction method for speaker identification using a classification algorithm. In AIP Conference Proceedings (Vol. 2656, No. 1, p. 020022). AIP Publishing LLC.

[7] *M. Abdullaeva, I. Khujayorov and M. Ochilov*. "Formant set as a main parameter for recognizing vowels of the Uzbek language," 2021 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2021, pp. 1-5, doi: 10.1109/ICISCT52966.2021.9670268.

[8] *Pande, Vinod & Kale, Vijay*. (2023). Speakers Identification Using Diarization Techniques. 10.2991/978-94-6463-136-4_80.

[9] *Gomez, Antonio*. (2022). Speaker Diarization and Identification from Single-Channel Classroom Audio Recording Using Virtual Microphones. 10.48550/arXiv.2207.00660.

[10] *Grozdić, Đorđe & Jovičić, Slobodan & Saric, Zoran & Subotić, Irina*. (2015). Comparison of GMM/UBM and i-vector based speaker recognition systems.

[11] *Pappagari, Raghavendra & Wang, Tianzi & Villalba, Jesús & Chen, Nanxin & Dehak, Najim*. (2020). x-vectors meet emotions: A study on dependencies between emotion and speaker recognition. 10.48550/arXiv.2002.05039.

[12] De*splanques, Brecht & Thienpondt, Jenthe & Demuynck, Kris*. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. 10.21437/Interspeech.2020-2650.

[13] *Chung, Joon Son & Nagrani, Arsha & Zisserman, Andrew*. (2018). VoxCeleb2: Deep Speaker Recognition. 1086-1090. 10.21437/Interspeech.2018-1929.

[14] *Wang, Qiongqiong & Lee, Kong Aik & Liu, Tianchi*. (2022). Scoring of Large-Margin Embeddings for Speaker Verification: Cosine or PLDA?. 10.48550/arXiv.2204.03965.
[15] *Kanagasundaram, Ahilan & Vogt, Robbie & Dean, David & Sridharan, Sridha*. (2012). PLDA based Speaker Recognition on Short Utterances.
[16] *K. Shukurov, U. Berdanov, U. Khasanov, S. Kholdorov and B. Turaev*. "The role of adaptive filters in the recognition of speech commands," 2021 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2021, pp. 1-4, doi: 10.1109/ICISCT52966.2021.9670084.
[17] *Saritha, Dr & Laskar, Mohammad & Kirupakaran, Anish & Laskar, Rabul & Choudhury, Madhuchhanda.* (2024). ReptoNet: A 3D Log Mel Spectrogram-Based Few-Shot Speaker Identification with Reptile Algorithm. Arabian Journal for Science and Engineering. 50. 10.1007/s13369-024-09426-3.
[18] *Dawalatabad, Nauman & Ravanelli, Mirco & Grondin, François & Thienpondt, Jenthe & Desplanques, Brecht & Na, Hwidong.* (2021). ECAPA-TDNN Embeddings for Speaker Diarization. 3560-3564. 10.21437/Interspeech.2021-941.
[19] *Garcia, Edel.* (2015). Cosine Similarity Tutorial.
[20] *Prince, Simon & Elder, James*. (2007). Probabilistic Linear Discriminant Analysis for Inferences About Identity. IEEE 11th International Conference on Computer Vision. 1-8. 10.1109/ICCV.2007.4409052.
[21] *Abdullaeva M.I., Juraev D.B., Ochilov M.M., Rakhimov M.F.* (2023). Uzbek Speech Synthesis Using Deep Learning Algorithms. In: Zaynidinov, H., Singh, M., Tiwary, U.S., Singh, D. (eds) Intelligent Human Computer Interaction. IHCI 2022. Lecture Notes in Computer Science, vol 13741. Springer, Cham. https://doi.org/10.1007/978-3-031-27199-1_5.
[22] *Ghosh, Debalina & Debnath, Depanwita & Bose, Saikat.* (2012). A Comparative Study of Performance of Fpga Based Mel Filter Bank & Bark Filter Bank. International Journal of Artificial Intelligence & Applications. 3.
[23] *Mukhamadiyev, A., Mukhiddinov, M., Khujayarov, I., Ochilov, M., & Cho, J.* (2023). Development of Language Models for Continuous Uzbek Speech Recognition System. Sensors, 23(3), 1145. https://doi.org/10.3390/s23031145
[24] *Jung, Youngmoon & Kim, Younggwan & Lim, Hyungjun & Kim, Hoirin.* (2017). Linear-scale filterbank for deep neural network-based voice activity detection. 1-5. 10.1109/ICSDA.2017.8384446.
[25] Sn*yder, David & Garcia-Romero, Daniel & Sell, Gregory & Povey, Daniel & Khudanpur, Sanjeev.* (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. 5329-5333. 10.1109/ICASSP.2018.8461375.
[26] *Loweimi, Erfan & Qian, Mengjie & Knill, Kate & Gales, M.J.F*. (2024). On the Usefulness of Speaker Embeddings for Speaker Retrieval in the Wild: A Comparative Study of x-vector and ECAPA-TDNN Models. 3774-3778. 10.21437/Interspeech.2024-161.
[27] *Berrar, Daniel.* (2018). Bayes' Theorem and Naive Bayes Classifier. 10.1016/B978-0-12-809633-8.20473-1.
[28] *Bahmaninezhad, Fahimeh & Hansen, John.* (2017). i-Vector/PLDA speaker recognition using support vectors with discriminant analysis. 10.1109/ICASSP.2017.7953190.

Поступила в редакцию 15.09.2025

Citation: *Shukurov K.E., Xasanov U.K.* (2025). Nutq signallari asosida so'zlovchilarni tanib olish algoritmlarining samaradorligini baholash. Raqamli texnologiyalarning nazariy va amaliy masalalari xalqaro jurnali. 9(1). – B. 80-89. https://doi.org/10.62132/ijdt.v9i1.325.

PERFORMANCE EVALUATION OF SPEAKER IDENTIFICATION ALGORITHMS USING SPEECH SIGNAL FEATURES

Shukurov K.E.[1], *Khasanov U.K.*[1]

[1] Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

Abstract. This article analyzes the effectiveness of using different models in speaker recognition processes and selects the best one for the system. In terms of accuracy and speed performance of the system, the classical MFCC + cosine similarity and modern x-vector, ECAPA-TDNN + PLDA architectures are compared. Based on the data set generated from

different speakers, the accuracy, f1-score, EER, latency, and GPU load indicators of the models are evaluated. According to the experimental results, the ECAPA-TDNN model outperforms the other models with an accuracy of 95.7%. Since the speaker recognition stage is also important for speaker separation systems, accuracy indicators are of high relevance. The ECAPA-TDNN + PLDA model offers good solutions in terms of using computational resources, working with large data sets, and analyzing their data.

Keywords: speaker identification, ECAPA-TDNN, x-vector, MFCC, log-Mel, PLDA, cosine similarity, deep learning, feature vector, speech biometrics, AM-softmax.

ОЦЕНКА ЭФФЕКТИВНОСТИ АЛГОРИТМОВ РАСПОЗНАВАНИЯ ДИКТОРОВ НА ОСНОВЕ РЕЧЕВЫХ СИГНАЛОВ

Шукуров К.Э.[1], *Хасанов У.К.*[1]

[1] Ташкентский университет информационных технологий имени Мухаммада аль-Хорезми, Ташкент, Узбекистан

Аннотация. В этой статье анализируется эффективность использования различных моделей в процессах распознавания говорящих и выбирается наилучшая для системы. С точки зрения точности и быстродействия системы сравниваются классическая архитектура MFCC + косинусное сходство и современная архитектура x-вектор, ECAPA-TDNN + PLDA. На основе набора данных, сгенерированного от разных дикторов, оцениваются показатели точности, f1-оценки, EER, задержки и загрузки графического процессора моделей. Согласно экспериментальным результатам, модель ECAPA-TDNN превосходит другие модели с точностью 95,7%. Поскольку этап распознавания говорящего также важен для систем разделения говорящих, показатели точности имеют высокую актуальность. Модель ECAPA-TDNN + PLDA предлагает хорошие решения с точки зрения использования вычислительных ресурсов, работы с большими наборами данных и их анализа.

Ключевые слова: идентификация говорящего, ECAPA-TDNN, x-вектор, MFCC, логарифмическое сходство, PLDA, косинусное сходство, глубокое обучение, вектор признаков, речевая биометрия, AM-softmax.