ISSN: 3030-3346

SO‘ZLOVCHILARNI AJRATISHDA ECAPA-TDNN ASOSIDAGI XUSUSIYATLAR VEKTORLARIDAN FOYDALANISH

Umidjon Xasanov Komiljon o‘g‘li

Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti, Sun’iy intellekt kafedrasi assistenti
Email. umidjon0923@gmai.com

K E Y W O R D S

So‘zlovchilarni ajratish, so‘zlovchilarni ifodalovchi xususiyatlar vektori, ECAPA-TDNN, X-vektor, Spektral klasterlash

A B S T R A C T

So‘zlovchilarni ajratishda xususiyatlar vektorlarini shakllantirish eng muhim bosqichlardan biri hisoblanadi. Zamonaviy chuqur neyron tarmoqlar so‘zlovchining individual xususiyatlarini aniqlash va ularni ifodalashda yuqori aniqlikni ta’minlay oladi. Xususan, x-vector kabi mashhur xususiyatlar vektorlari hozirgi tizimlarda asosiy komponent sifatida qo‘llanib kelmoqda. So‘nggi yillarda TDNN arxitekturasi ustida turli takomillashtirishlar ishlab chiqildi. ECAPA-TDNN modeli esa o‘ziga xos kanal e’tibor mexanizmi va ko‘p qatlamli agregatsiya orqali so‘zlovchini tekshirish vazifalarida yuqori samaradorlikni ko‘rsatdi. Tadqiqot mazkur modelni diarizatsiya sohasiga tatbiq etish imkoniyatlarini namoyish etadi.

KIRISH

So‘zlovchilarni ajratish texnologiyasi suhbat davomida “qaysi vaqtda kim nutq so‘zladi?” degan savolga aniq javob berishga xizmat qiladi[1,2]. Ushbu yondashuv sun’iy intellektga asoslangan muloqot tizimlarining muhim qismi bo‘lib, telefon suhbatlarini tahlil qilishdan tortib, televideniye yangiliklari, yig‘ilishlar va tibbiy yozuvlarni qayta ishlashgacha bo‘lgan ko‘plab sohalarda qo‘llaniladi. Zamonaviy diarizatsiya tizimlari asosan chuqur neyron tarmoqlar yordamida shakllantirilgan so‘zlovchi embeddinglariga tayanadi. Mazkur embeddinglar nutqdagi shaxsga xos xususiyatlarni ifodalaydi va keyinchalik samarali klasterlash algoritmlari orqali turli so‘zlovchilarni ajratish imkonini beradi[2].

So‘zlovchilarni ajratish bo‘yicha so‘nggi yillarda sezilarli taraqqiyot kuzatilgan bo‘lsa-da, bu yo‘nalish hanuzgacha nutqni qayta ishlashdagi eng murakkab masalalardan biri sanaladi[3]. Mazkur sohada olib borilayotgan tadqiqotlar juda faol bo‘lib, ular DIHARD kabi xalqaro tanlovlar orqali qo‘llab-quvvatlanmoqda[4]. Klasterlash jarayonida bir nechta yondashuvlar taklif qilingan: masalan, yuqoridan pastga yoki pastdan yuqoriga asoslangan agglomerativ klasterlash usullari.

So‘zlovchilarni ajratishda chuqur neyron tarmoqlardan olingan embeddinglar muhim rol o‘ynaydi. So‘nggi yillarda ishlab chiqilgan d-vector, c-vector va x-vector kabi modellar nutqda shaxsga xos akustik xususiyatlarni aniq ifodalash imkoniyatini yaratdi[5-7]. Ular orasida x-vector modeli alohida o‘rin tutadi: ushbu yondashuv Time Delay Neural Network (TDNN) arxitekturasiga asoslangan bo‘lib, bugungi kunda zamonaviy diarizatsiya tizimlarining eng ishonchli va keng qo‘llaniladigan tarkibiy qismi sifatida qaraladi[2]. Ushbu model o‘zining barqarorligi bilan boshqa uslublarga taqqoslaganda sezilarli afzalliklarga ega.

So‘zlovchilarni ajratish bilan bir qatorda so‘zlovchilarni aniqlash jarayoni nutqni qayta ishlash sohasining asosiy yo‘nalishlaridan biri. Ushbu jarayonning asosiy maqsadi berilgan nutq signalidan foydalanib, uni ma’lum bir shaxs bila bog‘lashdir. U uchun amaliy jihatidan ikki xil yondashuv qo‘llaniladi: identifikatsiya va verifikatsiya[8,9]. Identifikatsiya masalasida tizim mavjud ma’lumotlar bazasidagi barcha shaxslarni qidirib, qaysi shaxs gapirayotganini aniqlaydi. Verifikatsiya jarayonida esa tizim kiritilgan nutqni faqat bitta ma’lum shaxsga tegishliligini tasdiqlaydi yoki rad etadi.

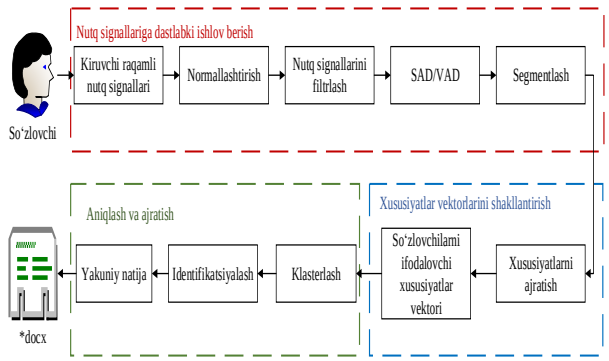
So‘zlovchilarni aniqlash texnologiyasi ko‘plab real amaliyotlarda o‘z o‘rniga ega. Masalan, xavfsizlik tizimlarida ovoz orqali shaxsni autentifikatsiya qilish, bank va moliyaviy

© *Umidjon Xasanov Komiljon o‘g‘li*

xizmatlarda foydalanuvchini aniqlash, call-markazlarda mijozlarni tezkor tanib olish, telemeditsina va masofaviy ta’lim tizimlarida shaxsiylashtirilgan xizmat ko‘rsatishni ta’minlash kabi vazifalarda samarali qo‘llaniladi[8].

ECAPA – TDNN ASOSIDA SO‘ZLOVCHILARNI AJRATISH TIZIMLARI

So‘nggi yillarda x-vector asosidagi TDNN arxitekturasi ustida turli takomillashtirishlar ishlab chiqildi. Ularning eng e’tiborlisi sifatida ECAPA-TDNN (ing: Emphasized Channel Attention, Propagation and Aggregation) modeli ko‘rsatib o‘tiladi. Mazkur arxitektura kanal va kontekstga mos e’tibor mexanizmi, ko‘p qatlamli xususiyatlarni birlashtirish (MFA), shuningdek Squeeze-Excitation (SE) va residual bloklarni o‘zida mujassamlashtiradi[8]. Buning natijasida model so‘zlovchini tekshirish (speaker verification) vazifalarida yuqori samaradorlikka erishadi va hatto Short-duration Speaker Verification (SdSV) musobaqasida eng yaxshi ko‘rsatkichlarni namoyish etgan. Bu esa qisqa segmentli nutqlarda ham aniq natija berishi sababli, uni so‘zlovchilarni ajratish vazifasida ham istiqbolli model sifatida qo‘llash imkonini beradi[9,10].



1-rasm. Nutq signallari orqali so‘zlovchilarni ajratish va aniqlash integratsion tizimining umumiy sxemasi

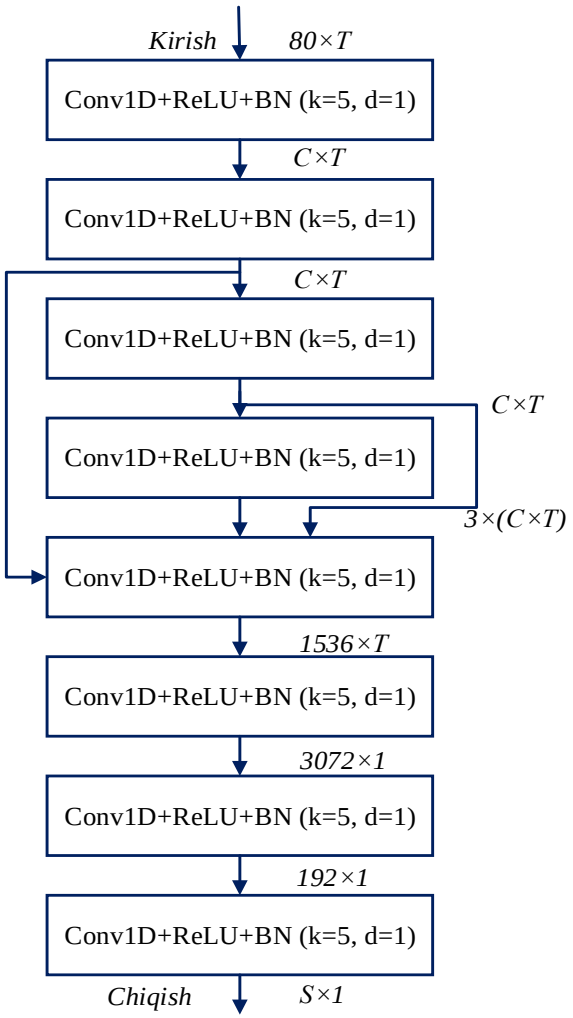
Olib boriladigan tadqiqotlarda e’tibor beradigan asosiy ikki yo‘nalishga qaratiladi. Birinchisi ECAPA-TDNN arxitekturasini so‘zlovchilarni ajratish va aniqlash vazifasiga tatbiq etiladi. So‘zlovchilarni tekshirish jarayonlarida yaxshi samara bergan ushbu model so‘zlovchilarni ajratish (diarizatsiya) jarayonlari uchun moslashtirish tizim samaradorligini sezilarli darajada oshiradi. Ikkinchidan esa ishlab chiqiladigan modelning turli xil sharoitlarda barqaror ishlashini ta’minlash maqsadida keng qamrovli nutq sifatini yaxshilash usullari qo‘llaniladi. Bunda nutq signallari turli xil ishlov berish jarayonlaridan o‘tkaziladi. Jumladan, nutq faolligi mavjud qismlarini aniqlash, jimlik sohalarini olib tashlash, tezlikni oshirish, shovqin va reverbaratsiyalarni yo‘qotish uchun moslashuvchan va WPE filtrlash algoritmlaridan tashkil topgan gibrid usulini amalga oshirish va boshqalar. Shu bilan birgalikda sun’iy hosil qilingan ma’lumotlar to‘plamidan foydalanilgan holda ma’lumotlar to‘plamlarini boyitish va ularga ishlov berish ishlab chiqiladigan modelni turli sharoitlarni qamrab olishini ko‘rsatadi.

So‘zlovchilarni ajratish jarayonlarida asosiy va muhim vazifalardan biri nutq signallaridan ishonchli va barqaror so‘zlovchini ifodalovchi xususiyatlar vektorilarini olishdir. Zamonaviy yondashuvlarda so‘zlovchilarni ifodalovchi xususiyatlar vektori chuqur neyron tarmoqlar orqali olinadi[10,11].

Tarmoq katta hajmdagi turli so‘zlovchilar nutqlari asosida o‘qitiladi va keyinchalik turli uzunlikdagi kirish segmentlari maxsus pooling qatlami yordamida bir xil o‘lchamdagi tasvirga keltiriladi. Shu tariqa hosil qilingan xususiyatlar vektorlar keyingi bosqichlarda so‘zlovchilarni ajratish va aniqlash uchun asos bo‘lib xizmat qiladi.

ECAPA-TDNN arxitekturasi o‘zining x-vector modeliga tayangan tuzilishi bilan ajralib turadi. Biroq, unda qo‘shimcha mexanizmlar joriy qilingan bo‘lib, ular embeddinglarning yanada mukammal shakllanishini ta’minlaydi[8,11]. Masalan, kanal va kontekstga bog‘liq e’tibor mexanizmi orqali tarmoq har bir freymni alohida qayta ishlash imkoniga ega bo‘ladi. Bundan tashqari, Squeeze-Excitation bloklari orqali oraliq qatlamlardan global kontekst ma’lumotlari olinadi[12], Res2-bloklar yordamida esa parametrlar soni kamaytirilib, hisoblash samaradorligi oshiriladi[13]. Yana bir muhim jihat ko‘p qatlamli xususiyatlarni birlashtirish (Multi-layer Feature Aggregation, MFA) bo‘lib, u turli qatlamlardan olingan xususiyatlarni statistik poolingdan oldin birlashtirib, yakuniy embeddingni boyitadi[14].

(cc) BY © Umidjon Xasanov Komiljon o‘g‘li



2-rasm. ECAPA-TDNN modelining blok sxemasi. C – kanal o‘lchami, T – xususiyat xaritalarining vaqt o‘lchami, k – konvolutsiya qatlamidagi yadro o‘lchami, d esa Conv1D yoki SE-Res2Block qatlamlaridagi dilatsiya oralig‘ini anglatadi.

O‘quv jarayonida model an’anaviy softmax o‘rniga AAM-softmax yo‘qotish funksiyasi bilan optimallashtiriladi. Ushbu yondashuv xususiyatlar vektorlari orasidagi kosinus masofani bevosita minimallashtirishga qaratilgan bo‘lib, natijada murakkab baholash tizimlari, masalan, PLDA ga ehtiyoj qolmaydi. Shu tariqa, tizim yanada ixcham va tezkor ishlash imkoniyatiga ega bo‘ladi[15].

Nutq signallarida turli shovqinlar, reverbaratsiya va akustik buzilishlar mavjudligi sababli modelni barqaror ishlashi juda muhim. Shuning uchun olib borilayotgan ishlarda keng qamrovli ma’lumotlarni ko‘paytirish usuli qo‘llaniladi. Asosiy farqi shundaki an’anaviy usullarda signalga bitta yoki bir nechta buzilishi qo‘shilib, keyin tarmoqqa o‘qitilsa, taklif etilayotgan modelda har bir nutq namunasi turli ko‘rinishda qayta ishlanib barchasi bitta o‘quv tarkibida birlashtiriladi[16-17]. Mazkur jarayonda bir nechta usullardan foydalaniladi.

1-jadval.

Nutq signalida qo‘llaniladigan ma’lumotlarni ko‘paytirish usullari

Usul nomi	Ta’rif
Wavefrom dropout	Signalning ayrim qismlari nol qiymatlari bilan almashtiriladi.
Frequency dropout	Chastota spektrining ba’zi diapazonlari filtrlab tashlanadi yoki to‘sib qo‘yiladi.
Tezlikni o‘zgartirish	Nutq signali biroz tezroq yoki sekinroq qilib qayta diskretlash orqali o‘zgartiriladi.
Rever	Signal tasodifiy xona impuls reaksiyasi bilan konvolutsiya qilinadi.
Shovqin qo‘shish	Nutqqa turli SNR darajalarida tashqi shovqin qo‘shiladi.
Gibrid buzilish	Shovqin va reverbaratsiya birgalikda qo‘llaniladi.

Bunday yondashuvning afzalligi shundaki, har bir o‘quv davrida model bitta jumlaning turli xil “versiyalari”ni ko‘radi. Bu esa neyron tarmoq parametrlarini turli akustik o‘zgarishlarga nisbatan mustahkam qiladi va umumiylashuv qobiliyatini kuchaytiradi. Natijada tizim real sharoitlarda, turli muhitlarda ham barqaror ishlash imkoniyatiga ega bo‘ladi.

Spektral klasterlash. So‘zlovchilarni ajratish tizimlarida so‘zlovchilarni ifodalovchi xususiyatlar vektorlarini guruhlash bosqichi muhim o‘rin tutadi. Ushbu bosqichda bir nechta yondashuvlar mavjud bo‘lsa-da, oxirgi yillarda spektral klasterlash usuli enf samarali usullardan biri sifatida taklif etiladi. Bu spektral usul an’anaviy aglomerativ iearxik klasterlashdan farqli ravishda, grafik nazariya tamoyillariga asoslanadi. Birinchi navbatda nutq segmentlari o‘rtasidagi o‘xshashlikni o‘lchash uchun affinlik matritsasi tuziladi. Bu jarayonda kosinus o‘xshashlik mezoni qo‘llaniladi. Matritsadagi kichik qiymatlar olib tashlanib, muhim bog‘lanishlarga e’tibor qaratiladi. Keyingi bosqichida esa ushbu ishlab chiqilgan affinlik matritsasi asosida Laplas matritsasini hosil qilinadi va unga xos qiymat-dekompozitsiyasi

(cc) BY © Umidjon Xasanov Komiljon o‘g‘li

qo‘llaniladi. Natijada asosiy eignvektorlar topilib, ular yangi fazodagi spektral tasvirlarni (so‘zlovchilarni ifodalovchi xususiyatlar vektorini) hosil qiladi.

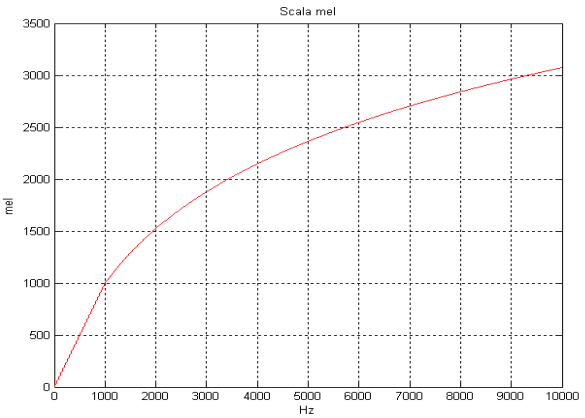
Spektral klasterlashning asosiy afzalliklaridan biri shundaki u nutq segmentlari sonini avtomatik aniqlash imkonini beradi. Buning uchun “eng katta xosvektor” me’zoni qo‘llaniladi, ya’ni ketma-ket xos qiymatlar orasidagi eng katta farq qaysi joyda kuzatilsa, u klasterlar sonini belgilash uchun mezon bo‘lib xizmat qiladi. Shundan so‘ng hosil bo‘lgan spektral embeddinglar odatiy k-means algoritmi yordamida guruhlarga ajratiladi.

Mazkur uslub murakkab akustik sharoitlarda ham samarali natija ko‘rsatadi. Ayniqsa, ko‘p so‘zlovchili suhbatlarda qisqa segmentlarni aniq ajratish imkoniyatini beradi. Shu sababli spektral klasterlash bugungi kunda zamonaviy diarizatsiya tizimlarida asosiy komponentlardan biri sifatida keng qo‘llanilmoqda.

ECAPA-TDNN UCHUN XUSUSIYATLARNI TANLASH

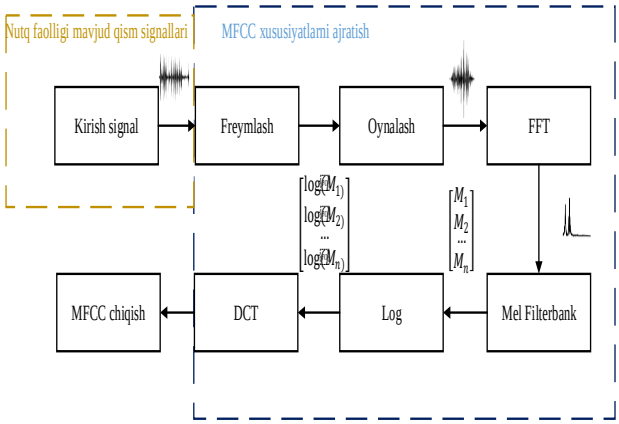
ECAPA-TDNN arxitekturasi samarali ishlashi uchun nutq signalidan olinadigan kirish xususiyatlari (features) hal qiluvchi ahamiyat kasb etadi. Odatda nutq signallarini qayta ishlash jarayonlarida xususiyatlarni ajratishda MFCC, PLP,CQCC, log-Mel kabi o‘nlab xususiyatlar ajratiladi. Bu xususiyatlar asosida nutq signallariga intellektual ishlov berish amalga oshiriladi[19].

MFCC texnologiyasi nutqni va so‘zlovchilarni tanib olish tizimlarida asosiy xususiyat sifatida qo‘llanib keladi. MFCC lar nutq signalining qisqa segmentlari bo‘yicha Mel chastota shkalasida spektral kuchlanishni hisoblab, uni diskret kosinus o‘zgartirish (DCT) orqali kepstral koeffitsientlarga o‘tkazadi. MFCC lar ma’lum darajada inson eshituv tizimini taqlid qiladi va signal spektrini ixcham ko‘rinishda, past o‘lchamli vektor (odatda 12-13 asosiy koeffitsient) sifatida beradi[20].



3-rasm. Mel shklasi

MFCC hisoblash uchun audio signal 20–25 ms uzunlikdagi oynalarga bo‘linadi (masalan, 25 ms) va har 10 ms da bitta freym olinadi. Har bir freymning quvvat spektri Mel-chastota shkalasidagi bank filtrlardan o‘tkazilib, logarifmik miqyosga o‘tkaziladi va so‘ngra DCT tatbiq qilinadi. Shunday qilib, MFCC spektrning umumiy shaklini siqilgan ko‘rinishda beradi hamda nutqning barqaror tafovutli xususiyatlarini (vokal trakting shakli) aks ettiradi



4-rasm. MFCC xususiyatlarini olish jarayoni

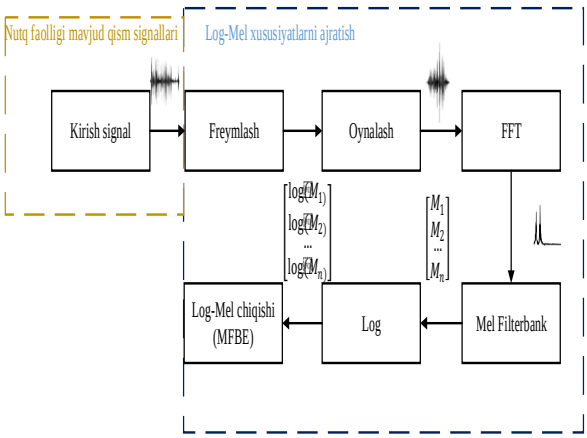
An’anaviy so‘zlovchilarni ajratish tizimlari so‘zlovchi ovozining tembr va artikulyatsion xususiyatlarini ifodalash uchun asosan MFCC kabi qisqa muddatli spektral xususiyatlardan foydalanadi. Biroq, MFCC xususiyatlari bevosita so‘zlovchini farqlash uchun optimallashtirilmagan, chunki ular nafaqat so‘zlovchilarga oid ma’lumotni, balki kanalning o‘zgarishlari, fon shovqinlari va gapirilayotgan fonemalarga bog‘liq xususiyatlarni ham o‘z ichiga oladi.

So‘nggi yillarda chuqur o‘rganish modellari rivojlanishi bilan log-Mel filtrbank xususiyatlariga qiziqish ortdi. Log-Mel spektrogrammalar bu

(cc) BY © Umidjon Xasanov Komiljon o‘g‘li

MFCC olinishi jarayonidagi oraliq mahsulot bo'lib, avvallari DCT orqali MFCC ga "siqib" olingan ma'lumotni to'liqroq saqlaydi. Dastlabki tadqiqotlar shuni ko'rsatdiki, katta miqdordagi ma'lumotlar va kuchli konvolyutsion neyrotarmoqlar yordamida to'g'ridan-to'g'ri mel-spektral tasvirlar ustida o'qitish yuqori natija beradi[20].

Log-Mel xususiyatlar esa MFCC olishdagi DCT bosqichigacha bo'lgan qiymatlar matritsasidir ya'ni logaritmik mel-spektrogramma. Bu matritsa odatda 40 dan 80 gacha o'lcham (mel-bandlar soni) ga ega bo'ladi (ECAPA-TDNNda 80 o'lchamli mel-filtrbank ishlatiladi) va har bir ustun audio freymning mel spektral kuchlanishlarini ifodalaydi. 80-o'lchamli log-Mel filtrbank har bir 25 ms li freym uchun olinib, model kirishiga beriladi. MFCC dan farqli o'laroq, log-Mel spektrogramma katta o'lchamli va o'zaro bog'liq komponentlarga ega, lekin unda nutq tovushining mayda spektral detallarigacha saqlanadi. Neyron tarmog'i konvolyutsion qatlamlar yordamida ushbu mel-spektragrammadagi naqshlarni o'zi mustaqil ravishda "o'rganadi", MFCC dagi kabi oldindan belgilangan ortogonal o'zgartirishga ehtiyoj yo'q[20].



5-rasm. Log-Mel xususiyatlarini olish jarayoni

ECAPA-TDNN ASOSIDA SO'ZLOVCHINI IFODALOVCHI XUSUSIYATLAR VEKTORLARINI SHAKLLANTIRISH

So'zlovchilarga xos xususiyatlar vektorlarini hosil qilish jarayoni bir nechta bosqichdan iborat(2-rasm). Birinchi navbatda kirish nutq signali $s(t)$ qisqa muddatli oynalarga bo'linadi. Oyna uzunligi 25 ms, hamda siljish 10 ms ni tashkil etadi. Har bir ajratilgan oyna uchun quvvat spektrlari hosil qilinadi.

$$X(\omega, n) = \left| \sum_{t=0}^{T-1} s(t + nH) \cdot \omega(t) e^{-i\omega t} \right|^2 \quad (1)$$

bu yerda:

$\omega(t)$ – Hamming oyna funksiyasi,

H – siljish.

Keyingi bosqichda Mel filtrlash qo'llaniladi va log-Mel energiyalari olinadi.

$$F_m(n) = log\left(\sum_{\omega} H_m(\omega) \cdot X(\omega, n) \right) \quad (2)$$

bu yerda:

$H_m(\omega)$ – Mel filtri.

Ushbu ish natijasida har bir fraym uchun M=80 o'lchamdagi log-Mel koeffisentli vektor olinadi.

ECAPA-TDNN arxitekturasida 1D-konvolutsiyalar vaqt bo'yicha o'zgarishlarni ushlaydi.

$$y_t^c = \sum_{i=1}^{k} W_i^c \cdot x_{t-i \cdot d} + b^c \quad (3)$$

bu yerda:

- k – kernel o'lchami,
- d – dilatsiya parametri,
- W_i^c – og'irliklar,
- x_t – kirish xususiyatlari.

Residual va SE-bloklar yordamida kanallararo o'zaro bog'liqlik hisobga olinadi. Endi ko'p qatlamli xususiyatlarni birlashtirish amalga oshiriladi. Buning uchun turli qatlamlardan chiqqan belgilar umumlashtiriladi.

$$h_t = Concat\left(h_t^{(1)}, h_t^{(2)}, \ldots, h_t^{(L)}\right) \quad (4)$$

bu yerda:

$h_t^{(i)}$ – i-qatlamdan olingan xususiyatlar.

Olingan vaqt bo'yicha xususiyatlar bir vektor ko'rinishiga keltiriladi. Kirish sifatida $[\mu, \sigma]$ qo'shib olinadi. Bu jarayonni statistik yig'ish (pooling) deb nomlanadi.

$$\mu = \frac{1}{T} \sum_{t=1}^{T} h_t, \quad \sigma = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (h_t - \mu)^2} \quad (5)$$

© Umidjon Xasanov Komiljon o'g'li

Yig'ish jarayonidan keyin olingan vektor bir nechta to'liq bog'langan qatlamlardan o'tkazilib yakuniy so'zlovchini ifodalovchi xususiyatlar vektorini e shakllanadi.

$$e = f_\theta([\mu, \sigma]) \quad (6)$$

bu yerda:

f_θ – neyron tarmoq parametrlar to'plami.

Modelni o'qitishda klassik softmax o'rniga Additive Angular Margin Softmax (AAM-Softmax) qo'llaniladi. Bu usul vektorlarni kosinus masofa bo'yicha ajratishni kuchaytiradi.

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

bu yerda:

- s – masshtablash parametri,
- m – burchakli margin,
- θ_{y_i} – haqiqiy sinf bilan burchak.

Hosil bo'ladigan so'zlovchini ifodalovchi xususiyat vektori $e \in R^d$ (ko'p hollarda 192, 256) keyingi klasterlash va identifatsiya bosqichlari uchun foydalaniladi.

TAJRIBALAR NATIJALARI

ECAPA-TDNN modeliga kirish xususiyati sifatida 80-o'lchamli log-Mel filterbank energiyalari tanlandi. Har bir segment bo'yicha o'rtacha qiymat normallashtirilib, neyron tarmoqqa uzatildi. Model parametrlari Adam optimizatori yordamida yangilanib, o'qitishda siklik o'qitish tezligi (Cyclical Learning Rate, CLR) bilan qo'llanildi. MFCC koeffisentlaridan ko'ra log-Mel energiya koeffisentlari tanib olish masalalarida yaxshi natijalarini ko'rsatdi. Buni aniqlash uchun Voxceleb xalqaro nutq korpusidan hamda avval ishlab chiqilgan o'zbek tili uchun yaratilgan nutq korpusidan foydalanildi. Baholash mezonlari sifatida DER (Diarization Error Rate) – so'zlovchini to'g'ri ajratish xatoligi, EER (Equal Error Rate) – so'zlovchini to'g'ri taxmin qilish xatoligi, JI (Jackard Index) – anotatsiyalarning mosligi, aniqlik (accuracy) ko'rsatkichlari olindi.

Xususiyat turi	Diarizatsiya (DER)	Identifikatsiya (EER)	Identifikatsiya (aniqlik)	Diarizatsiya (JI)
MFCC	15.4%	9.8%	83.4-85.1%	90.23%
Log-Mel	9.8%	5.6%	89.3-90.15%	93.40%

Ko'rinib turibdiki, log-Mel xususiyatlari har ikki vazifada ham MFCCga nisbatan yaxshiroq natijalarga olib kelmoqda. Xususan, diarizatsiyada DER ko'rsatkichini ~1.5 barobarga, identifikatsiyada esa EER ko'rsatkichini ~2 barobarga yaxshilashga erishilmoqda. Aniqlik bo'yicha ham log-Mel foydasiga ~6-7% absolyut farq kuzatilgan.

Bundan tashqari turli klasterlash algoritmlari yordami so'zlovchilarni ajratishni amalga oshirishda ECAPA-TDNN modeli hamda an'anviy qo'llab kelingan X-vector modelini qo'llab natijlarni solishtirildi. Natijalarni baholash uchun esa DER hamda JI mezonlari baholandi.

Clustering algorithms	Model	DER	JI
K-means	MFCC, X-vector	25.61	89.14
Spectral		22.65	88.13
Agglomerative		24.65	90.15
DBSCAN		27.65	87.23
K-means	Log - Mel, ECAPA-TDNN	17.81	90.84
Spectral		9.81	93.40
Agglomerative		10.24	92.45
DBSCAN		11.02	89.14

Jadval natijalari shuni ko'rsatadiki xususiyatlar va so'zlovchini ifodalovchi xususiyatlar vektori tanlovi so'zlovchilarni ajratish sifatiga sezilarli darajada ta'sir ko'rsatadi.

XULOSA

Olib borilgan tadqiqotlar shuni ko'rsatmoqdaki, ECAPA-TDNN arxitekturasi asosida ishlab chiqilgan so'zlovchilarni ajratish modeli mustahkam so'zlovchi ifodalovchi xususiyatla vektorini hosil qiladi, shu bilan birgalikda klasterlash bosqichlarida ham ishonchli samara beradi. Bundan tashqari spektral klasterlash kombinatsiyasi tizimning umumiy aniqligini oshiradi.Xususiyatlar tanlovi amalga oshirilganda so'zlovchini ajratish jarayonlarida Log-Mel+ECAPA-TDNN kombinatsiyasi MFCC+X-vector dan ancha samarali. Zamonaviy so'zlovchilarni ajratish tizimlarida ECAPA-TDNN modeli asosida hosil qilingan so'zlovchini ifodalovchi xususiyatlar vektorlari hamda spektral klasterlash algoritmlari yaxshi samarali yechim sifatida qaraladi.

FOYDALANILGAN ADABIYOTLAR

1. X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent

(cc) BY © Umidjon Xasanov Komiljon o'g'li

research," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 2, pp. 356–370, 2012.
2. T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," 2021, arXiv:2101.09624.
3. G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge," in Proc. Interspeech, 2018, pp. 2808–2812.
4. N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The second DIHARD diarization challenge: Dataset, task, and baselines," in Proc. Interspeech, 2019, pp. 978– 982.
5. L. Wan, Q. Wang, A. Papir, and I. Lopez-Moreno, "Generalized end-to-end loss for speaker verification," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4879–4883, 2018.
6. G. Sun, C. Zhang, and P. C. Woodland, "Speaker diarisation using 2d self-attentive combination of embeddings," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5801–5805, 2019.
7. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5329–5333, 2018.
8. B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPATDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in Proc. Interspeech, 2020, pp. 3830–3834.
9. H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "SdSV challenge 2020: Large-scale evaluation of short-duration speaker verification," in Proc. Interspeech, 2020, pp. 731–735.
10. J. Thienpondt, B. Desplanques, and K. Demuynck, "Cross-lingual speaker verification with domain-balanced hard prototype mining and language-dependent score normalization," in Proc. Interspeech, 2020, pp. 756–760.
11. D. Garcia-Romero, G. Sell, and A. McCree, "Magneto: X-vector magnitude estimation network plus offset for improved speaker recognition," in Proc. Odyssey, 2020, pp. 1–8.
12. J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in IEEE/CVF CVPR, 2018, pp. 7132–7141.
13. S. Gao, M.-M. Cheng, K. Zhao, X. Zhang, M.-H. Yang, and P. H. S. Torr, "Res2Net: A new multi-scale backbone architecture," IEEE TPAMI, pp. 652–662, 2019.
14. Z. Gao, Y. Song, I. McLoughlin, P. Li, Y. Jiang, and L.-R. Dai, "Improving aggregation and loss function for better embedding learning in end-to-end speaker verification system," in Proc. Interspeech, 2019, pp. 361–365.
15. J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in IEEE/CVF CVPR, 2019, pp. 4685–4694.
16. D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in Proc. Interspeech, 2019, pp. 2613–2617.
17. N. Dawalatabad, S. Madikeri, C. C. Sekhar, and H. A. Murthy, "Novel architectures for unsupervised information bottleneck based speaker diarization of meetings," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 14–27, 2021.
18. E. Hoffer, T. Ben-Nun, I. Hubara, N. Giladi, T. Hoefler, and D. Soudry, "Augment your batch: better training with larger batches," CoRR, vol. abs/1901.09335, 2019.
19. Shukurov K.E., Kaxxarov A.A., Xasanov U.K., "O'qituvchisiz o'qitish asosida so'zlovchilarni ajratish", "Raqamli transformatsiya: axborot texnologiyalari, sun'iy intellekt va iqtisodiyotda yangi davr", 442-447b, 2025
20. Xasanov U.K., To'rayev X.Sh., Tolibova N.D., "So'zlovchilarni ajratish jarayonlarida nutq signallariga dastlabki ishlov berish", 153-157b, 2025

© Umidjon Xasanov Komiljon o'g'li