1 Asmens atpažinimaspagal balsą

1.1 Atpažinimo pagal balsą uždaviniai

1 paveikslėlis. Asmens atpažinimo pagal kalbą algoritmų klasifikacija

Asmens identifikavimpo ir verifikavimo uždaviniai Asmens balso verifikacijos atveju atliekamas palyginimas vienas su vienu, t. y. asmuo pareiškia savo tapatybę, pasako kalbos pavyzdį ir tuomet jo balso pavyzdys palyginimas su anksčiau įrašytais to asmens kalbos pavyzdžiais. Jei pasakyto kalbos pavyzdžio panašumas į anksčiau pasakytus kalbos pavyzdžius viršija pasirinktą slenkstį, asmens tapatybė patvirtinama (verifikuojama); priešingu atveju asmens tapatybė nepatvirtinama (asmuo neverifikuojamas). Verifikacijos uždavinys natūraliai kyla ribotos prieigos sistemose, pasienio kontrolės punktuose. Kiekvienas verifikacijos algoritmas daro dviejų rūšių klaidas: tam tikras autentiškų balsų yra neverifikuojamas ir tam tikra dalis apsišaukėlių balsų yra patvirtinama autentiškais. Pirmoji klaida išreikšta procentais žymima FRR (angl. *False Rejection Rate*), o antroji FAR (angl. *False Acceptance Rate*). Šių dviejų klaidų procentinė reikšmės priklauso nuo pasirinkto sprendimo slenksčio reikšmės; kuo didesnis panašumo reikšmių slenkstis, tuo didesnis FRR ir mažesnis FAR, ir, atvirkščiai, kuo mažesnis slenkstis, tuo mažesnis FRR ir didesnis FAR. Parametrinės kreivės.

$$x(t) = FAR(t),$$

$$y(t) = FRR(t),$$

kur parametras t yra panšumo reikšmių slenkstis vadinamas DET (*Detection Error Tradeoff*) kreive ir vizualiai įvertina diktoriaus balso verifikacijos algoritmo kokybę. Kuo DET kreivės grafikas žemesnis tuo verifikacijos algoritmo kokybė geresnė.

Identifikacijos atveju atliekamas vienas su daugeliu arba vienas su N palyginimas. Šiuo atveju asmuo neprisistato ir balso identifikavimo sistema turi rasti ir pateikti tolimesnei analizei labiausiai panašius į duoto kalbėtojo šnekos pavyzdžio balsus iš turimos diktorių balsų pavyzdžių duomenų Balso identifikavimo uždavinys kyla kriminalistikoje, kai, pavyzdžiui, teismas duoda sankciją pasiklausyti asmens pokalbių ir reikia automatiškai identifikuoti kokio asmens kalbos įrašas užfiksuotas. Teismuose taip pat užsakomos asmens balso ekspertizės, kurių vienas tikslų identifikuoti kokių asmenų balsai girdimi teismui pateiktuose įrašuose. Identifikuojant asmenį pagal jo kalbos pavyzdį, yra apskaičiuojami pateikto identifikavimui balso pavyzdžio panašumai iki visų turimų asmenų balsų pavyzdžių ir gauti panašumai surūšiuojami pateikiant vartotojui labiausiai tikėtinų balsų sąrašą. Didžiausio panašumo balsų pora pateikia hipotezę apie kalbėtojos tapatybę. Egzistuoja du identifikavimo uždavinio porūšiai: atviras ir uždaras. Atviro uždavinio atveju nėra žinoma ar pateiktas kalbos pavyzdys iš vis priklauso kokiam nors turimos duomenų bazės asmeniui ir pradžioje reikia nuspręsti ar balso pavyzdys priklauso kokiam nors duomenų bazėje esančiam asmeniui ir, jei priklauso, rasti tą asmenį. Uždaros aibės atveju iš anksto žinoma, kad pateiktas identifikavimui balsas priklauso duomenų bazėje esančiam kokiam nors asmeniui ir identifikavimo metu reikia rasti asmenį, kurio balsas panašiausias į pateiktam identifikavimui balsą. Uždaros aibės uždavinys yra lengvesnis, nes šiuo atveju asmuo, kurio balso panašumas didžiausias į pateikto kalbos pavyzdžio panašumą, gali būti identifikuotas kalbėtoju, o atviros aibės atveju reikia dar papildomai nuspręsti ar didžiausio panašumo balsas yra to paties asmens ar ne. Verifikacijos uždavinys yra atskiras atviros aivės identifikavimo atvejis, kai $N = 1$.

Asmens identifikavimo algoritmų kokybė vertinama eiliškumo (angl. *ranking*) kreive. Piešiant šią kreivę abscisių ašyje atidedamas skaičius nuo 1 iki N, o ordinačių ašyje atidedamas kaupiamasis procentas kokia dalis buvo identifikuota x ašies vietoje ir mažiau.. Pavyzdžiui, tarkime N=5 ir identifikacijai buvo pateikta 20 kalbos pavyzdžių. Tarkime, remiantis balsų porų panašumu, 11 kalbos pavyzdžių buvo teisingai identifikuota pirmoje vietoje, 4 antroje, 3 trečioje , 0 ketvirtoje ir likę 2 benktoje vietoje. Tuomet šio kalbančiųjų balsų identifikavimo uždavinio eiliškumo kreivę sudarys taškai

$$(1, 55), (2, 75), (3, 90), (4, 90), (5, 100).$$

Priklausantis ir nepriklausantis nuo teksto asmens kalbos identifikavimas Kitas kriterijus pagal kurį skirstomis asmens balso identifikavimo uždaviniai, yra sakomo teksto tipas. Jei sakomas tekstas žinomas iš anksto ir yra tas pats prisistatymo ir duomenų bazės įraše, tokiu atveju sakoma, kad asmens balso identifikavimo algoritmas priklausantis nuo teksto. Priklausantis nuo

teksto asmens balso identifikavimo sistemos reikalauja žymiai trumpesnių rbalso pavyzdžių trukmių. Pavyzdžiui gana patikimam asmens balso identifikavimui gali pakakti trumpos frazės, pvz. Aš esu Jonas Jonaitis, mano tabelio numeris 123. Kita algoritmų rūšis identifikuoja asmenį nepriklausomai nuo sakomo teksto. Šiuo atveju kalbos pavyzdžio trukmė turėtų būti trijų-penkių min. Taip yra todėl, kad tik tariant pakankamai ilgai tekstą susikaupia pakankamai patikima ir stabili balso atskirų elementų statistika, kuri leidžia patikimai atpažinti kalbantįjį pagal jo balsą. Nepriklausomų nuo teksto algoritmų privalumas, kad jai paremtas asmens identifikavimo sistemas sunkiau apeiti. Pavyzdžiui priklausančią nuo tariamo teksto sistemą nesunku apeiti įsirašius iš anksto žinomą kalbos pavyzdį ir jį įgarsinant prisistatymo metu. Yra ir tarpinis variantas, kai sakomas tekstas nėra iš anksto žinomas, tačiau visų tariamų atskirų žodžių pavyzdžiai turimi kalbos pavyzdžių duomenų bazėje. Tipinis pavyzdys: įrašų bazėje saugomi visų galimų dešimtainių skaitmenų tarimo pavyzdžiai, o prisistatymo metu prašoma pasakyti iš anksto nežinomo daugiaženklio skaičiaus pavyzdį. Tokia sistema prisistatymo metu taip pat nereikalauja ilgų kalbos pavyzdžių ir ją sunku apeiti su iš anksto pasiruoštais įrašais.

Priklausančių ir nepriklausančių nuo teksto asmens balso identifikavimo algoritmai naudoja skirtingas technikas. Priklausančio nuo teksto atveju populiarūs tokie metodai:

1. Laiko mastelio keitimo, DTW (angl. *Dynamic Time Warping*), (Rabiner at all, 1978, White, Neely, 1976), technika. Ši technika elegantiškai išsprendžia skirtingu greičiu sakomų to paties teksto.
2. Paslėptojo Markovo modelio, HMM (angl. *Hidden Markov Model*) [Rabiner, Juang, 1986]

Nepriklausomose nuo teksto asmens balso identifikavimo sistemose dominuoja tokios technikos:

1. Gauso mišinio modelis GMM (Gaussian Mixture Model) (Reynolds, 1995)
2. Vektorinio Kvantavimo (VQ, Matsui, Furui, 1992)
3. Aritmetinė Harmoninė Sferiškumo metrika (Harmonic Sphericity measure, AHS, Bimbot, Mathan, 1993)
4. Įvairios paslėptojo Markovo modelio variacijos (Hidden Markov Model, HMM) [Rabiner, Juang, 1986]

Nepriklausomos nuo teksto sistemos turi papildomą informatyvumą, nes kiekvienas kalbėtojas turi individualią dažniausiai naudojamų žodžių statistiką, kurią įvertinus galima panaudoti kaip papildomą informaciją identifikuojant asmenį.

1.2 Kalbos pirminis apdorojimas

Pradžioje rekomenduojama paryškino kalbos signalo aukštus dažnius. Tai atliekama pritaikant pradiniam kalbos signalui $x = x(t)$ tokį filtrą:

$$y(t) = x(t) - ax(t-1).$$

Filtro parametras a imamas iš intervalo [0.95, 0.99] ir priklauso nuo skaitmeninio kalbos signalo imčių dažnio. Kai kurie autoriau siūlo adaptuoti parametrą a kiekvienam apdorojamam kalbos kadrui parenkamas adaptyviai. Dažniausiai a reikšmė parenkama eksperimentiniu būdu maksimizuojand asmens kalbos identifikavimo ir verifikavimo kokybę. Jei šio parametro reikšmė mažai gerina atpažinimo kokybę, siūloma iš vis praleisti šią filtravimo procedūrą.

Pradinis kalbos signalas dalinamas į *kadrus*, kuriems pritaikoma lango funkcija, kad sumažinti kraštinių trūkių signalo reikšmių įtaką apskaičiuojamiems kalbos kadro požymiams. Kiekvieno kadro trukmė yra 20-30 milisekundžių. Kad padidinti kadrų kiekį ir padaryti gretimų kadrų požymius mažiau trūkius, naudojamas gretimų kadrų persidengimas ir dažniausiai laiko trukmė tarp gretimų kadrų pasirenkama 10 milisekundžių. Šios kadrų trukmės patikrintos empiriškai ir pagrįsto žmogaus kalbos dinamikos įverčiais. Kadangi per sekundę pasakomas apie dešimt fonetinių vienetų turintis signalas, tai, siekiams kadre paimto signalo stacionarumo, jo trukmė neturi viršyti $1/10 = 0.1$ sek. Kadangi dalis fonetinių vienetų gali būti kelis kartus trumpesni už vidurkį, tai rekomenduojama vieno kadro trukmė 20-30 milisekundžių. Tokios trukmės signalo spektro įverčio skiriamoji geba 100 hercų. Toliau mažinant kadro trukmę mažėtų kadro spektro skiriamoji geba, kas yra nenaudinga, nes vyrų kalbos pagrindinio tono (balso stygų virpėjimo dažnis) reikšmė svyruoja apie 100 hercų. Lango funkcija dažniausiai pasirenkama Hammingo arba Hanningo. Abu langai sumažina kraštines kalbos kadro reikšmes, o tai padidina signalo/triukšmo santykį dažnių srityje. Greitoji Furjė transformacija (FFT, 1965)-(FFT,1989) atspindi kalbos kadro spektrinę sudėtį. Kad atlikti diskrečiąją Furjė transformaciją greitai, kadro imčių skaičius padidinamas iki artimiausio dvejeto sveikojo laipsnio papildant signalą kraštuose reikiamu kiekiu nuliais.

Reziumuojant, išvardinsime naudojamas pirminio kalbos signalo apdorojimo procedūras.

- Detektuojami ir eliminuojami tylos ir nedidelės energijos kalbos fragmentai.
- Kalbos sigalas paryškinamas naudojant pirmos eilė filtrą $1 - 0.95z$.
- Kalbos įrašas suskaidomas į 30 msek. trukmės fragmentus (kadrus) naudojant 20 msek. kadrų persidengimą.
- Kiekvieno kadro imtys apdorotos naudojant Hanningo langą.

1.3 Melų skalės Kepstras

FFT modulio kvadratas išreiškia kalbos kadro galingumo spektrą. FFT galingumo spektro įvertis labai nereguliarus, todėl dažnai naudojami juostiniai kalbos signalo filtrai, kurių centrinio dažnio reikšmė ir juostos plotis charakterizuoja filtrą. Juostinį spektro filtrą galima parametrizuoti naudojant filtro kairįjį, centrinį ir dešinįjį dažnius. Filtras gali būti trikampiu ar eksponentiniu.. Imituojant žmogaus klausos charakteristikas, juostinių filtrų centriniai dažniai keičiami panaudojant Bark arba Melo skalę. Melo skalėje trikampių juostinių filtrų centriniai dažniai apibrėžiami tokia taisykle (Fant and Gunnar, 1968):

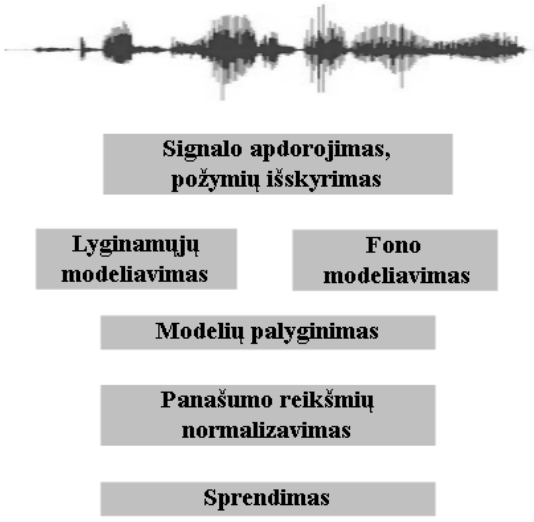
$$f_{Mel} = 1000 \log_2(1 + \frac{f}{1000}).$$

Egzistuoja sudėtingesni Melo skalės dažnių apibrėžimai, bet jų visų esmė ta pati: iki 1000 hercų centriniai dažniai keičiami beveik pastoviu žingsniu, o didesniems dažniams pradeda logaritmiškai didėti. Toliau imamas FFT galingumo spektro logaritmas, dauginamas iš 20, kad gauti spektro reikšmes decibelais (dB), ir dauginamas iš centruoto trikampio lango įvertinant svertinio spektro reikšmę. Tokiu būdu kiekvienas kalbos kadras aprašomas požymių vektoriumi, kurio dimensija priklauso nuo trikampių filtrų kiekio, o kiekviena komponentė yra spektro logaritminio galingumo tam tikroje dažnių juostoje svertinė vertė. Tačiau tokie požymiai turi perteklinės informacijos, todėl atliekama papildoma transformacija. Dažniausiai ta transformacija yra diskrečioje kosinusų transformacija, kurios rezultatas vadinamas kepstro koeficientais: (Bogert at all, 1963), (Oppenheim and R.W. Schafer, 1968):

$$c_n = \sum_{k=1}^{K} S_k \cos(\frac{(n-0.5)(k-0.5)\pi}{K}), \quad n = 1, 2, \ldots, N.$$

Čia K yra Melo skalės spektro galingumo logaritmų reikšmių S_k kiekis, o $N \leq K$yra kepstro koeficientų skaičius.

2 Kalbos modeliavimas



2 paveikslėlis. Tipinė automatinio asmens atpažinimo pagal balsą schema

2 pav. pavaizduota bendra kalbančiojo atpažinimo schema. Aprašydami kiekvieną schemos modelį rėmėmės (Bimbot at all, 2004) šaltiniu.

2.1 Tiesinė prognozė

Poliniame *tiesinės prognozės* (angl. *Linear Prediction* (LP)) modelyje kalbos signalo reikėmės x_n aproksimuojamos ankstesnių reikšmių tiesiniu dariniu (Itakura and Saito, 1968):

$$x_n = -\sum_{n=1}^{P} a_k x_{n-k} + Ge_n,$$

kur P yra LP modelio eilė, a_k tiesinės prognozės koeficientai (angl. *linear prediction coefficients, LPC*), G žadinimo signalo stiprumas ir e_n normuotos energijos šaltinio signalas. LPC parametrai surandami minimizuojant aproksimacijos

$$\hat{x}_n = -\sum_{k=1}^{P} a_k x_{n-k}$$

vidutinę kvadratinę paklaidą. Kaip taisyklė šaltinio reikšmės e_n yra nemodeliuojamos. Tai nėra idealus sprendimas, nes šaltinio signale yra informacija apie pagrindinį toną, kuris naudingas identifikuojant kalbantįjį. Tai galima kompensuoti tiesiogiai įvertinant vokalizuotų kalbos fragmentų pagrindinį toną. LP modelio spektras apibrėžiamas formule

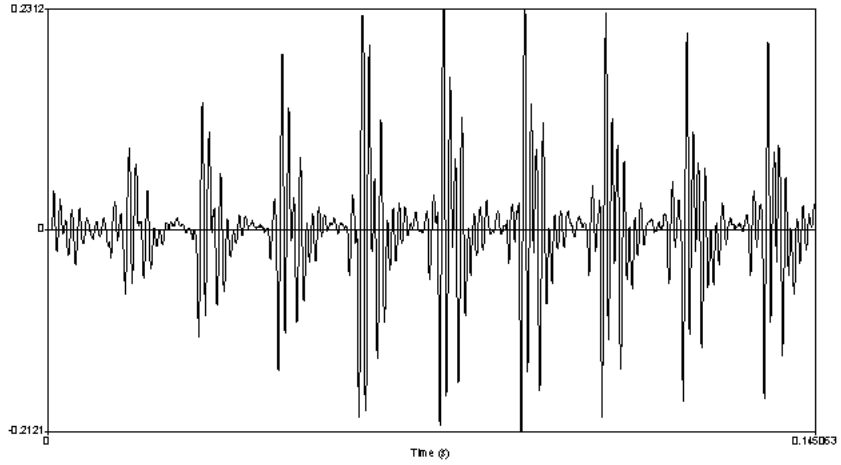
$$H(z) = \frac{G}{1 + \sum_{k=1}^{P} a_k z^{-k}} = \frac{G}{A(z)},$$

kur $A(z)$ atvirkštinis P-osios LP polinio modelio filras. Minimizuojant vidutinę paklaidų $x_n - \hat{x}_n = \epsilon_n$ kvadratinę paklaidą, tiesinės prognozės koeficientams gaunama tiesinių lygčių sistema, kuri lengvai ir greitai išsprendžiama kompiuterio pagalba.

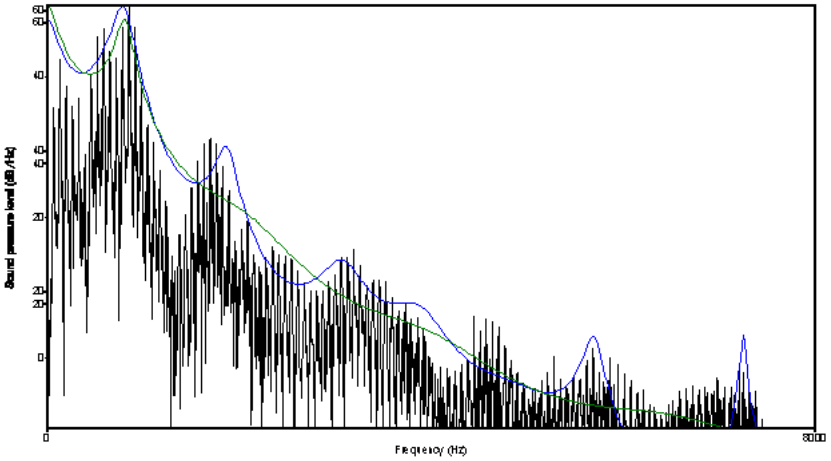
3 paveikslėlis iliustruoja Lietuviško garso a skaitmeninį įrašą, o 4 - jo spektro įvertį gautą trimis skirtingais metodais. Juoda kreivė žymi FFT galingumo spektro logaritmą, žalia - 8-osios eilės LP modeliu gautą to paties garso spektro įvertintą, o mėlyna - to paties LP modelio, kurio eilė du kartus didesnė ($P = 16$). Bendrai FFT spektrų įverčiai pasižymi perdėtu detalumu ir yra nestabilūs mažiems kalbos signalo požyčiams ar kadro pradžios poslinkiui. LP modeliu gautas spektro įvertis yra FFT spektro įverčio tam tikra gaubiamoji, kurios detalumas priklauso nuo modelio eilės P..

2.2 LP modelio kepstro koeficientai

LP modelio koeficientai apskaičiuojami kiekvienam kalbos fragmentui. Tiesiogiai asmens identifikavui tiesinės prognozės koeficientai naudojami retai. Taip yra dėl to, kad a_k parametrai nestabilūs mažiems signalo pokyčiams ir neturi skaidrios fizikinės interpretacijos. Yra žinoma, kad kepstro koeficientus galima įvertinti tiesiniais prognozės koeficientų dariniais. Jei modelio eilė P artėja į begalybę, aproksimacijos virsta lygybėmis (LPCC, 1977).. Tiesinės išraiškos, kurios



3 paveikslėlis. Lietuviškos fonemos a skaitmeninis įrašas



4 paveikslėlis. a fonemos spektro skirtingi įverčiai

konvertuoja LPC koeficientus į LP *kepstrinius koeficientus* (LPC į LPCC) yra tokios:

$$
\begin{aligned}
c_0 &= \ln G, \\
c_m &= a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}, \quad 1 \le m \le P, \\
c_m &= \sum_{k=m-P}^{m-1} \frac{k}{m} c_k a_{m-k}, \quad m > P.
\end{aligned}
$$

Papildomi požymiai Melo skalės arba LP kepstro koeficientai suteikia galimybę lengvai sumažinti kanalo įtaką. Stacionarūs kanalo iškraipymai apytiksliai įtakoja signalą kaip tam tikras filtras. Kadangi kepstro koeficientai įvertinami panaudojant galingumo spektro logaritmą, kanalo filtras pavirsta adityviu nariu. Kadangi šis adityvus narys apytiksliai pastovus visiems kadrams, jį galima eliminuoti apskaičiuojant visų kadrų kepstro koeficientų vidurkį ir jį atimant iš kiekvieno kadro kepstro koeficientų. Ši operacija vadinama kepstro vidurkio atėmimu (angl. *cepstral mean subtraction, CMS*). CMS operacija ženkliai padidina asmens balso identifikavimo kokybę kai yra stacionarūs kanalo iškraipymai. Tokie iškraipymai atsiranda kai lyginamųjų ir tiriamųjų balsai įrašomi skirtingomis priemonėmis (pavyzdžiui skirtingais mikrofonais, skirtingose aplinkose, pvz. kambarys vs. lauko sąlygos). Tačiau CMS technika nesumažina adityvaus triukšmo.

Kepstro koeficientai nesuteikia informaciją apie kalbos signalo dinamiką. Tuo tikslu siūloma naudoti gretimų kadrų kepstro koeficientų pirmosios ir antrosios eilės skirtumai Δ ir $\Delta\Delta$. Šie parametrai aproksimuoja pirmosios ir antrosios eilės kepstro koeficientų išvestines (Furui, 1981). Šios išvestinės įvertinamos tokiomis formulėmis:

$$
\begin{aligned}
\Delta c^m &= \frac{\sum_{k=-l}^{l} k c_{m+k}}{\sum_{k=-l}^{l} |k|}, \\
\Delta\Delta c^m &= \frac{\sum_{k=-l}^{l} k^2 c_{m+k}}{\sum_{k=-l}^{l} k^2},
\end{aligned}
$$

kur c su viršutiniu indeksu m žymi m-ojo kadro kepstro koeficientų vektorių, o parametras $l = 1, 2$ arba 3.. Pirmoji kepstro koeficientė komponentė ($c_= \ln G$) yra neinvariantiška įrašymo sąlygoms (garsumui) ir todėl neįtraukiama į požymių vektorių, tačiau Δ ir $\Delta\Delta$ nejautrūs įrašymo sąlygų garsumui ir todėl juos galima įtraukti į bendrą kadro požymių vektorių, kuris naudojamas asmens balso identifikavimui.

3 Kalbančiųjų modeliai ir jų palyginimas

When speech utterance is represented as a sequence of feature vectors it is called that features of the signal are extracted. To have possibility to compare extracted features the same type of features are selected and for target and for investigative speech examples. However different utterances have different textual content, have different duration and therefore directly frame-by-frame comparison of features can not be done. In this section we will provide an short introduction to features matching techniques. There are known two groups of measures that are used for estimation of speech utterances. The first group construct a statistical model for measured features vectors. If features $\mathbf{f}$ are K dimensional vectors a density function $d = d(\mathbf{f})$ that maximizes likelihood of observed features of the frames is constructed. If at authorisation process is observed speech frame with features vector $\mathbf{f}$, direct substitution of $\mathbf{f}$ to $d(\mathbf{f})$ gives likelihood of that frame for the target speaker with the density function $d = d(\mathbf{f})$. Such substitutions should be done for each frame and an average $d(\mathbf{f})$ value represents similarity measure of the two speakers models. Much quicker comparison of the two voices can be done by constructing density function and for investigative voice and estimating probability that two densities correspond to the random source of features vector $\mathbf{f}$. Another type of measures directly compares pairs of features vectors that correspond to different frame of the target and investigative voice and a global measure of similarity is constructed from local comparisons of similarity of pairs of frames. This technique is called template matching, is more intuitive, and as rule, is more expansive. The both types of measures have their merits and demerits, and therefore their combination is often used.

3.1 Šablonų modeliai

In simplest template model only a single template $\mathbf{f}$, which is the model target speech, is used. Template $\mathbf{f}$ belong to the linear space of all possible features vectors an can be defined as mean vector of features vectors of speech frames. Such choice minimizes mean square euclidean distance error between a fixed template and the all frames features vectors. If we have $\mathbf{f}^m,\ m = 1, 2, \ldots, M$, features vectors of the M frames of a target voice, than target speaker template would be

$$\bar{\mathbf{f}} = \frac{\sum_{m=1}^{M} \mathbf{f}^m}{M}.$$

Distance between feature vector $\mathbf{f}^m$ of a investigative m-th frame and target model $\bar{\mathbf{f}}$ is expressed by

$$d(\mathbf{f}^m, \bar{\mathbf{f}}) = \sqrt{(\mathbf{f}^m - \bar{\mathbf{f}})^T \mathbf{W} (\mathbf{f}^m - \bar{\mathbf{f}})}.$$

Here $\mathbf{W}$is a features components weighting matrix. Euclidean distance is defined by identity matrix, covariance matrix of frame features vectors define Mahalanobis distance. If initial features vectors are transformed to the space that basis

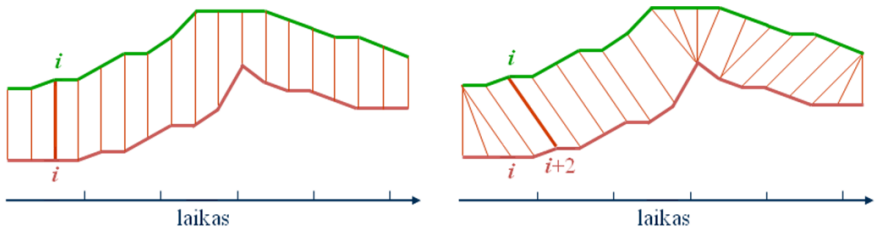
consist of orthogonal eigenvectors of the covariance matrix, the Mahalanobis distance is equal to Euclidean distance and computational cost the latter is much smaller (is proportional to the dimensionality of features vector) (Duda and Hart, 1973).

3.2 Dinaminis laiko mastelio keitimas (DTW)

If speaker recognition is text-dependent or text-prompted with vocabulary covered in saved utterances data-base, template matching is an intuitive and often used in speaker recognition. The idea is that even the sane text examples spoken by the same person are more similar than the ones of different speakers. The voice recognition becomes more easer if different speakers cooperate with authorization system and pronounce personalised utterances as I am Jonas Jonaitis, engineer, my personal number 375 and I am Petras Petraitis, my job position is in support division, personal number 781. It is naturally expect that average value of frame-to-frame distance of both utterances of the same text should be good discriminative characteristic for recognition of the claimed speaker. However in text-dependent and text-prompted case we will face by small variations in speed by which utterances are spoken. Dynamic Time Warping (DTW), (Rabiner at all, 1978) gives an elegant solution which in some sense optimally arranges the frames that should by paired in comparison of two utterances. The cost of DTW algorithm is moderate since in general distances between the all frames of two utterances should be estimated that consist qudratic complexity. Let suppose we have $\{\mathbf{f}^1, \mathbf{f}^2, \ldots, \mathbf{f}^M\}$ investigative voice features vectors and $\{\bar{\mathbf{f}}^1, \bar{\mathbf{f}}^2, \ldots, \bar{\mathbf{f}}^M\}$ target voice features vectors. Than DTW algorithm gives non-decreasing set of indexes $j(1), j(2), \ldots, j(L) \in \{1, 2, \ldots, M\}$ that minimizes with some additional conditions the average distance

$$d(\mathbf{f}, \bar{\mathbf{f}}) = \frac{\sum_{m=1}^{M} d(\mathbf{f}^m, \bar{\mathbf{f}}^{j(m)})}{M}.$$

Figure 3 illustrates identical alignment j(m) = m of frames of two curves that



5 paveikslėlis. Kalbos kadrų atitiktis be išlyginimo (kairiau) ir su DTW išlyginimu (dešiniau)

have the same number of points (left part) and the one which minimizes average

distance between two curves (right part). There are known some attempts to explore DTW method for text-independent speaker recognition. However DTW algorithm has quadratic complexity and text-independent speech examples are much longer that that weights direct application of DTW technique in such conditions.

3.3 Kalbos modelių palyginimas vektorinio kvantavimo metodu

The main drawback of DTW templates matching approach is that this technique does not work for text-independent speaker recognition. A direct on templates matching of two speech samples would be estimation of distances or similarities between all possible pairs of features vectors that correspond to two speech utterances and minimization of the obtained distances matrix by columns and rows and calculation of average minimal distances. However such direct approach would lead to big computational cost. For example if we have two utterances of 3 and 5 min. length and distance between neighbour frames is 10 milliseconds, the total number frame pairs similarity of which should be estimated will be 3x60x10 x 5x60x10 = 54 e 6 that sufficiently big number even for modern computers. Vector Quantization is an old well known technique which allows to reduce initial number of vectors by rounding them to centroids that consist so called codebook (VQ, 1987). Vectors of codebook are formed usually by some clustering procedure. Size of the codebook ranges in speaker recognition from 32 to 2048 and has tendency to grow in last years. Let C denotes the codebook constructed for target speaker vectors $\mathbf{f}^l$, Than average quantization distance of investigative voice features vectors $\mathbf{f}^m$defines distance between the two speakers. Formally for the distance we have such expression:

$$d(\mathbf{f}, \bar{\mathbf{f}} = \sum_{m=1}^{M} \min_{\bar{\mathbf{f}}} \frac{d(\mathbf{f}^m, \bar{\mathbf{f}})}{M}.$$

The vector quantization technique reduces computational costs and is often used as one of similarity/distance measure for voices comparison. To increase speed of comparison of two voices the features vectors of investigative voice can be vector quantized too and than can be used distances or similarities between codewords of the two vector codebooks however such approach decreases quality of speaker recognition. Sometimes such double quantization approach is used for initial selection or most similar pairs of voices that are further investigated by traditional Vector Quantization modeling.

3.4 Artimiausiojo kaimyno metodas

Nearest neighbours (NN) method combines strength of DTW and VQ methods. Unlike the VQ method, NN method keeps all features vectors of the target data (NN, 1993). For each test session frame is found the most similar enrolled target frame and inversely for each enrolled target frame is found the most similar test

session frame and the two series of minimal distances are averaged, This method is computationally most costly however gives the best results in recognition of text-independent speakers when the recognition is done by templates matching methodology.

3.5 Stochastiniai modeliai

Templates methods work well for text-dependent speaker recognition however are expansive and not state of the art quality when one needs text-independent recognition. In stochastic approach a density function is constructed which maximizes likelihood to observe features vectors that are observed for target speakers. For each target speaker a separate density function is constructed. Then estimating likelihood to observe features vectors of unknown speaker for all target models gives measure of probability that the unknown speaker has identity o a target speaker. So we have set of conditional probability distribution functions with the number of conditions equal to the number of target speakers. Conditional probability density function (pdf) of a target speaker is estimated from the set of training features vectors and can be parametric on non-parametric. In any case (parametric or non-parametric pdf) probability that features vectors of unknown speaker are generated by the claimed target model can be estimated. This probability gives unnormalized matching scores. To build parametric model, a specific form of pdf should be assumed and then the free parameters of the model are determined by maximization of likelihood of observed training features vectors. One possible assumption may be that the pdf is the multivariate normal density function. Then the free parameters of the model would be mean vector μ and covariance matrix C of the multivariate normal distribution. In this case value

$$p(\mathbf{f}^m|\text{ lyginamojo modelis }) = \sqrt{2\pi}^{-K}|C|^{-1/2}\exp(-\frac{(\mathbf{f}-\mu)^T C^{-1}(\mathbf{f}-\mu)}{2}).$$

Here K is dimension of frame features vector, —C— is determinant of the covariance matrix. Having features training vectors $\vec{\mathbf{f}}^l, l = 1,\ldots,L$, mean vector and covariance matrix of target model can be estimated by the following expressions:

$$\mu = \frac{\sum_{l=1}^{L}\vec{\mathbf{f}}^l}{L},$$
$$C = \frac{\sum_{l=1}^{L}(\vec{\mathbf{f}}^l-\mu)\cdot(\vec{\mathbf{f}}^l-\mu)^T}{L-1}.$$

Here . denotes point-wise multiplication. However multivariate normal distribution is too simplified approximation of real training vectors and therefore often is used Gaussian Mixture Model (GMM) in which density function is normalized sum of a few different multivariate normal distributions. We will give more detailed description of this model later. Although strictly speaking speech frames do

not provide independent features vectors it is assumed its independence that allows to estimate conditional probability of unknown speaker simply multiplying frames probabilities. Another very popular stochastic model is Hidden Markov Model (HMM) (Rabiner, Juang, 1986). Hidden Markov Model is double embedded stochastic process in the sense that the stochastic process is not directly observable. The HMM is defined by

1. finite set of states 2. NxN matrix of transition probabilities , that means transit at next time moment to the state j if we were at state i at current time. It is assumed that transition probabilities do not depend on time. 3. finite set of M observable symbols , 4. NxM matrix of probabilities , that means probability to observe symbol at state, 5. N probabilities that define state probabilities at initial moment.

Having observations set and HMM it is easy to calculate probability of such observation. However in practice HMM should be constructed from observations. For fixed parameter N the rest of HMM parameter and sequence of states are chosen by maximizing probability to have the observations set under the model and the states sequence. The two problems are solved using Baum-Welch and Viterbi algorithms (Juang, Rabiner, 1991)

3.6 Gauso mišinio modelis

The most popular stochastic model that long time is successfully applied for speaker recognition is Gaussian Mixture Model (GMM). The authors of this method are Reynolds and Rose (Reynolds, Rose, 1995). In this model pdf function is modelled by the expression:

$$p(\mathbf{f}^m|\text{ lyginamojo modelis }) = \sum_{i=1}^{I} p_i g_i(\mathbf{f}),$$

kur

$$g_i(\mathbf{f}) = \sqrt{2\pi}^{-K} |C_i|^{-1/2} \exp(-\frac{(\mathbf{f}-\mu_i)^T C_i^{-1}(\mathbf{f}-\mu_i)}{2})$$

are shifted multivariate normal distribution and

$$p_i \geq 0,\ i = 1, \ldots, I,\ \sum_{i=1}^{I} p_i = 1$$

weight of the shifted and scaled normal distributions. The complete Gaussian mixture density has I mean K dimensional vectors, KxK covariance matrices and positive weights. However it is assumed often that covariance matrices have simple structure, for example are diagonal, that save required for model memory and simplifies estimation of the model. GMM model has simple interpretation. Speech signals are composed by different phonemes that can by clustered in features space and each component of GMM density can represent a particular phoneme and the weights of mixture represents frequency/probability of occurrence of that phoneme. Mean vectors define acoustic positions of the phonemes and covariance

matrices C_i sharpness of localization of phonemes around their acoustic centre. GMM has advantage over VQ approach since the latter can be interpreted as an approximation of pdf by a discrete histogram with centers in codewords. On the over hand codewords of VQ can be used for initial positions of mean vectors that are later tuned by iteration process that maximizes a posteriori probability to observe training features vectors. Let $\lambda = (p_i, \mu_i, C_i),\ i = 1, 2, \ldots, I$, represents parameters of the GMM. Than having target training features vectors $\bar{\mathbf{f}}^l,\ l = 1, 2, \ldots, I$ the GMM parameters are found by maximizing the a posteriori probability

$$p(\bar{\mathbf{f}}|\lambda) = \prod_{l=1}^{L} p(\bar{\mathbf{f}}^l|\lambda).$$

The a posteriori probability highly non-linearly depends on the model parameters that forces to apply some iterative process for maximization of the probability. Having constructed GM target model the measure of correspondence of unknown voice to the target voice is simply estimated by

$$p(\mathbf{f}|\lambda) = \prod_{m=M}^{L} p(\mathbf{f}^m|\lambda),$$

where $\mathbf{f}^m,\ m = 1, \ldots, M$,are features vectors of unknown speaker voice utterance.

4 Grupinės delsos požymiai

4.1 LPC spektro fazės panaudojimas kalbančiojo identifikamui

Yra įprasta kalbos ir kalbančiojo atpažine naudoti šnekos fragmentų spektrinį tankį. Kepstro koeficientai ir formančių pozicijos įvertinamos naudojant spektrinį tankį, kuris išreiškiamas vien spektro moduliu nenaudojant fazės. Toks požiūris suformuotas gerai žinomais rezultatais apie žmogaus kalbos suvokimo specifiką eliminuoti informaciją apie signalo spektro fazę. Tačiau neaišku kuo remiantis dažnai ignoruojama informacija apie kalbos fragmento perdavimo funkcijos fazę. Mes siūlome LPC spektro fazę panaudoti kalbančiojo požymiams aprašyti ir pritaikyti ją automatiniam balso identifikavimui. Požymius sudaro perdavimo funkcijos gupinės delsos (group delay) pagrindu įvertinti kalbos atskirų kadrų poliai. Poros pirmoji komponentė pažymi grupinės delsos maksimumo argumentą, o antroji yra atsumo iki vienetinio apskritimo įvertis. Viso viename kalbos kadre yra įvertinama iki 7-ių grupinės delsos maksimumų. Pasiūlyta dviejų kadrų panašumo įvertinimo metrika, kuri remiasi aprašytais požymiais. Metrika yra optimizuota nuo teksto nepriklausomam diktoriaus atpažinimui darant prielaidą, kad tiriamame kalbos įraše gali būti keli kalbantieji. Atlikti tyrimai parodė, kad pasiūlyti kalbos požymiai patikimai atskiria skirtingus kalbėtojus ir gali būti sėkmingai kombinuojami su Mel keptro, formančių ir antiformančių ir pagrindinio tono požymiais.

4.2 Tiesinės prognozės modelis

Autoregresijos pagrindu apskaičiuotame tiesinės prognozės (LPC, angl. *Linear Prediction*) modelyje kalbos signalo yra eliminuota. Todėl pagal LPC modelį įvertinta perdavimo funkcija

$$s(z) = \frac{g}{\sum_{p=0}^{P} a_p z^p}, \quad |z| = 1,$$

Kiekvienam z suteikia informaciją apie perdavimo funkcijos amplitudę ir fazę. kuri neytakota pradinio kalbos signalo spektro fazė

Kalbančiojo identifikavimo automatiniai atpažinimo algoritmai dar nėra labai aukštos kokybės lyginant su biometriniais identifikavimo algoritmais grįstais pirštų ar delnų atspaudų, rainelių ar veidų pateikiama informacija. Mūsų galva kalbos analizėje yra nepakankamai įvertinama informacija, kurią supeikia klbos perdavimo funkcijos fazė. Tradiciniai galingumo spektru grįsta kalbos fragmentus aprašantys požymiai (formantės, kepstro coeficientai) tradiciškai yra įvertinami išnaudojant vien perdavimo funkcijos amplitudę. Mes siūlome alternatyvą - analogiškus požymius įvertinti remiantis vien perdavimo funkcijos faze. Kad išvengti kalbos kadro trakto parametrų stabilumo problemų, naudojame tradicinė tiesinės prognozės (LP - Linear Prediction) modelį. Derindami [1] ir [3] darbuose pateiktas technikas įvertiname LP modelio perdavimo funkcijos fazės požymius. [1] LP modelio spectro fazės trečiosios eilės išvestinės naudojamos įvertinti kalbos kadro trakto formantes. Mes naudoja naudojame tik pirmos ir antros eilės LP modelio perdavimo funkcios fazės išvestines. Antrosios eilės išvestinės nulio kirtimai suteikia informaciją apie formančių pozicijas, o pirmosios eilės išvestinės reikšmė maksimumo taške suteikia informaciją apie formantės stiprumą. [3] darbe parodytas ryšys tarp LP modelio perdavimo funkcijos ir linijinio spektro dažnių (LSF - Line Spectrum Frequencies). Tai padėjo mums atrasti simetrizuotą LP modelio perdavimo funkciojos modelio fazės išraišką kas taupo skaičiavimų laiką ir pasiūlo LPC spektro polių aproksimacijų formule. Gauso mišinio modelis (GMM - Gaussian Misture Model) (žiūr. [4]) vis dažniau naudojamas kalbos trakto požymių modeliavimui ir jų palyginimui. Kadangi mūsų požymių reikšmės apribotos stačiakampyje $(0, \pi) \times (0, 1)$, požymių skirstinių įvertinimui naudojome histogramų techniką ir pasiūlėme informacijos teorija grįsta dviejų kalbos traktų palyginimo metriką.

4.3 Tiesinė prognozė

Tiesinė prognozės (LP) modelyje [5] kalbos kadro imtys išreiškiamos forma

$$x_n = \sum_{i=1}^{P} a_i x_{n-i} + G e_n, \tag{1}$$

kur $a_1, a_2, \ldots, a_P$ yra Tiesinės Prognozės Koeficientai (LPC), P - modelio eilė, G yra šaltinio žadinimas, o e_n tiesinio modelio paklaidos. LPC modelio parametrai a_p yra įvertinami minimizuojant 30 ms. trukmės kadro aproksimacijos paklaidų

suminę energiją. Paprastumo dėlei laikysime, kad LP modelio eilė yra nelyginė, t.y. $P = 2M - 1$. z srityje LP modelis (1) atrodo taip

$$X(z) = GE(z)/A(z), \tag{2}$$

kur

$$A(z) = 1 - \sum_{i=1}^{2M-1} a_i z^{-i} \tag{3}$$

yra atvirkštinė perdavimo funkcija. Tisesioginė funkcija

$$H(z) = \frac{G}{A(z)} \tag{4}$$

kartais vadinama *LPC spektru* arba LP filtro *perdavimo funkcija* ir yra kalbos kadro spektro gaubiamoji, kurios detalumoas priklauso nuo modelio eilės P.

4.4 LP modelio fazė

Simetrinį ir antisimetrinį daugianarį $p(z)$ ir $q(z)$ formulėmis

$$p(z) = \frac{z^M A(z) + z^{-M} A(z^{-1})}{2}, \tag{5}$$

$$q(z) = \frac{z^M A(z) - z^{-M} A(z^{-1})}{2i}, \quad i = \sqrt{-1}. \tag{6}$$

$p(z)$ ir $q(z)$ daugianariai susieti su linijino spectro dažnių (LSF) simetriniu ir antisimetriniu daugianariu $P(z)$ ir $Q(z)$ tokiais sąryšiais

$$P(z) = A(z) + z^{-2M} A(z^{-1}) = 2z^{-M} p(z), \tag{7}$$

$$Q(z) = A(z) - z^{-2M} A(z^{-1}) = 2iz^{-M} q(z). \tag{8}$$

Vienetiniame apskritime $|z| = 1$ daugianariai $p(z)$ ir $q(z)$ yra įgyja realias reikšmes,

$$|A(z)|^2 = p(z)^2 + q(z)^2, \tag{9}$$

ir

$$p(z) + q(z)i = z^M A(z). \tag{10}$$

(9) ir (10) lygty parodo, kad perdavimo funkcijos dažnių atsakas ir fazė tenkina lygtis

$$|H(z)| = \frac{G}{\sqrt{p(z)^2 + q(z)^2}}, \tag{11}$$

ir

$$(\arg H)(e^{i\omega}) = \Phi(\omega) = M\omega - \arctan(\frac{q(e^{i\omega})}{p(e^{i\omega})}), \quad \omega \in [0, 2\pi). \tag{12}$$

4.5 LPC spektro fazės požymiai

LPC spektras gali būti išreikštas polių modelyje taip:

$$H(z) = \frac{G}{\prod_{m=1}^{P}(1 - r_m e^{i\alpha_m} z^{-1})}, \tag{13}$$

kur $r_m e^{i\alpha_m}$ yra LPC spektro m-ojo polio spindulys, o $\alpha_m \in [0, 2\pi)$ - polio kampinis dažnis. Iš (13) išplaukiam kad m-asis polis tiesiškai įtakoja LPC spektro fazę adityviu nariu

$$\arctan(\frac{r_m \sin(\omega - \alpha_m)}{1 - r_m \cos(\omega - \alpha_m)}).$$

Todėl fazės pirmoji ir antoji išvestinė tenkina tokias lygtis:

$$\frac{d\Phi(\omega)}{d\omega} = \sum_m \frac{r_m(\cos(\omega - \alpha_m) - r_m)}{1 - 2r_m \cos(\omega - \alpha_m) + r_m^2} \tag{14}$$

ir

$$\frac{d^2\Phi(\omega)}{d\omega^2} = -\sum_m \frac{r_m(1 - r_m^2)\sin(\omega - \alpha_m)}{(1 - 2r_m \cos(\omega - \alpha_m) + r_m^2)^2}. \tag{15}$$

Kad supaprasti ir pagreitinti skaičiavimus, polių tikslios vertės neapskaičiuojamos ir išvestinės (14) ir (15) yra surandamas skaitmeniškai diferiancijuojant (12) išraišką.

Iš (14) matome, kad stipriems poliams, kurių modulis r_m arti 1, galima tikėtis lokalaus LPC spektro fazės ekstremumo ω_m, kuris yra artimas kampiniam dažniui α_m. Lokalus maksimumas ω_m yra fazės antrosios eilės išvestinės nulio kirtimo taškas artimas α_m reikšmei. Naudojant (14) gauname

$$\Phi'(\omega_m) \approx \frac{r_m}{1 - r_m} \tag{16}$$

ir

$$r_m \approx \frac{\Phi'(\omega_m)}{1 + \Phi'(\omega_m)}. \tag{17}$$

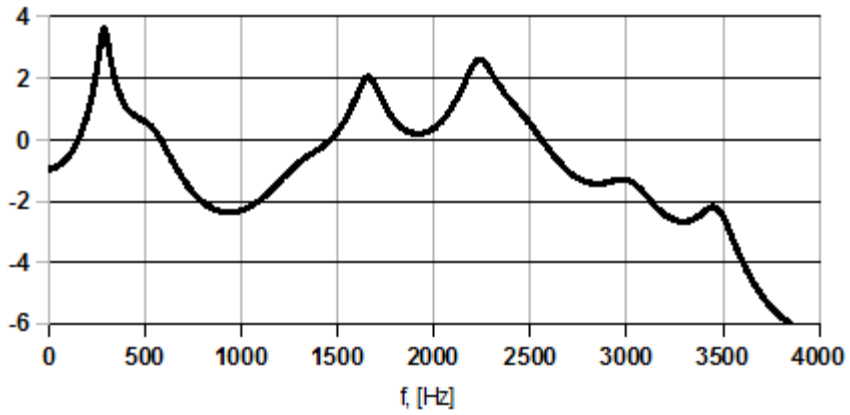
Todėl kalbos kadro LPC spektro fazės požymius apibrėžėme aibe skaičių porų

$$(\omega_m, \frac{1}{1 + \Phi'(\omega_m)}) = (\omega_m, \delta_m), \tag{18}$$

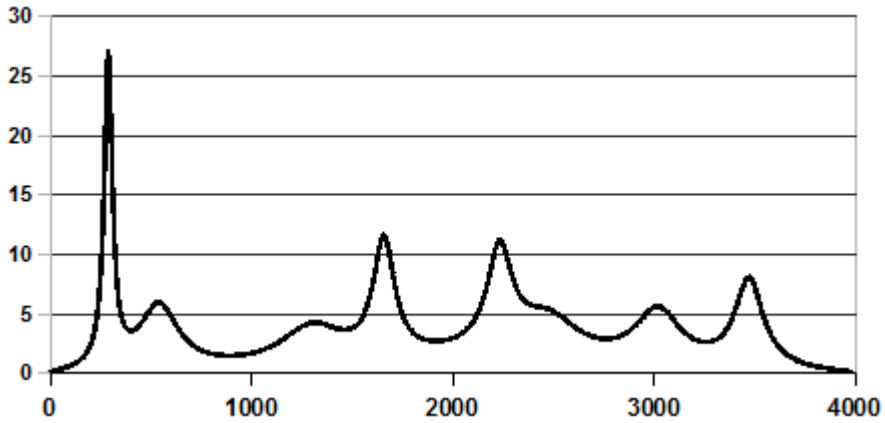
kur $\{\omega_m\}_m$ yra LPC spektro fazės antrosios eilės nulio kirtimai, kurie atitinka fazės kampinio greičio kitimo lokalaus maksimumo taškus priklausančius intervalui $(0, \pi)$ ir antru požymio poros skaičiumi

$$\delta_m = 1 - \frac{\Phi'(\omega_m)}{1 + \Phi'(\omega_m)} = \frac{1}{1 + \Phi'(\omega_m)} \tag{19}$$

įvertinamas spektro formantės plotis. Pažymėsime, kad mokslinėje literatūroje spektro fazės kitimo kampinis greitis dažnai vadinamas *grupine dels (group delay)*.

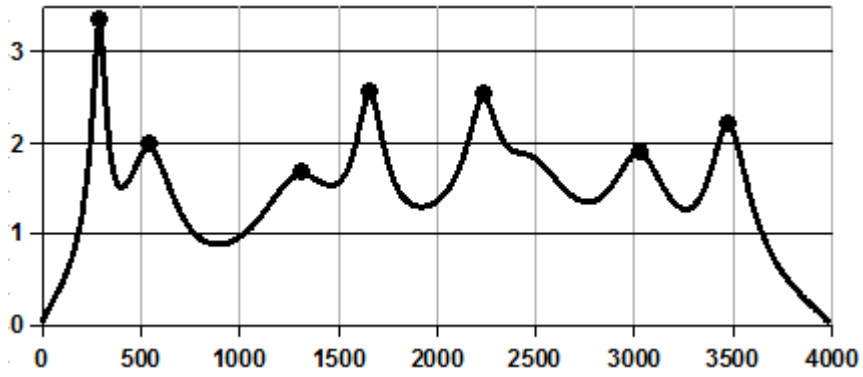


6 paveikslėlis. Kalbos kadro LPC spektro galingumo logaritmas



7 paveikslėlis. LPC spektro fazės pirmosios eilės išvestinė

Pav. 6 iliustruoja vieno kalbos kadro LPC galios spektro logaritmą, o pav. 7 vaizduoja to paties kadro LPC spektro fazės išvestinę. Naudojant (17) aproksimaciją galima visiems dažniams f įvertinti poliaus atstumą iki vienetinio apskritimo. 8 brėžinys vaizduoja atstumo įverčio log-aritmą. Brėžinyje paryškinti taškai pažymi kadro spektro poymius, kuriuos mes naudojame diktoriaus balsui atpažinti.



8 paveikslėlis. Spektro pločio logaritmas su pažymėtais vieno kalbos trakto kadro požymių taškais $(f_m, -log(\delta_m))$

5 Kalbos pavyzdžių panašumo metrika naudojama balsui atpažinti

Tarkime turime du skaitmeninės kalbos pavyzdžius $\{x_n\}$ ir $\{y_n\}$, kurių panašumą turime įvertinti. Tarkime $\{x_n\}$ imtys yra lyginamųjų aibės X atstovo, o $\{y_n\} = Y$ yra kito kalbos įrašo imtys, turios priklauso vienam, dviem ar daugiau tiriamųjų. Balso panašumo metriką turi įvertinti tikimybę, kad lyginamųjų atstovo X balsas skamba tiriamame Y įraše. Toks balso atpažinimo uždavinys natūraliai iškyla teismo balso pavyzdžių ekspertizėje, kai reikia atsakyti į klausimą ar duotame įraše Y skamba X asmens balsas. Teisminėje ekspertizėje lyginamajo balso pavyzdžiai gali būti įrašyti atskirame kanale arba rankiniu būdu išskirti iš daugiakanalio įrašo, o tiriamieji įrašai Y sudaro natūralius dažnai kelių asmenų kalbos įrašusu.

5.1 Požymių statistika

Pereitame skyrelyje įvedėme LP modeliu įvertintos perdavimo funkcijos fazės požymius, kurie yra aprašo grupinės delsos ekstremumus. Pažymėkime k-ojo kalbos kadro Grupinės Delsos (GD) ekstremumus (f_m^k, δ_m^k), kur f_m^k yra k-ojo

kadro m-ojo maksimumo dažnis ir δ_m^k m-ojo poliaus atstumo iki vienetinio apskritimo aproksimacija. Kalbos įrašas yra dalinamas į 1 sek. trukmės intervalus ir įvertinas spektro fazės požymių (f_m^k, δ_m^k) skirstinys. Kadango atstumas tarp gretimų kalbos kadrų 0.01 sek., vienos sekundės intervale yra apie $100(M-1)$ porų (f_m^k, δ_m^k). $(f_m^k, \delta_m^k) \in (0, \frac{FS}{2}) \times (0,1)$ skirstinys įvertinamas dalinant $(0, \frac{FS}{2}) \times (0,1)$ stačiakampį į $N \times L$ stačiakampių dalių ir apskaičiuojant kiek porų (f_m^k, δ_m^k) patenka į kiekvieną stačiakampį. Deformacijos parametras $\lambda = \lambda(FS)$ yra adaptuojamas įmčių dažniui FS kad padalinimas dažnių intervalo $(0, \frac{FS}{2})$ lygiomis dalimis apytiksliai atitinka Barko dažnių skalę. Galimų atstumų intervalas $(0,1)$ padalinamas augančia 10-ies Fibonači intervalų seka.

5.2 Vienos sekundės trukmės dviejų kalbos intervalų palyginimas

Trumpų vienos sekundės kalbos intervalų palyginimui naudojame informacijos metriką. Mes naudojame panašumo metriką kuri nepanašiems segmentams priskiria artimas nuliui reikšmes, o didėjant lyginamų fragmentų panašumui panašumo reikšmės didėja. Panašumo metrika apibrėžiama normuota dvirjų lyginamų segmentų tarpusavio informacija. Tegul $I = N \times L$ yra bendras padalinimo stačiakampių skaičius, $\{B_i\}_{i=1}^I$ - padalinimo stačiakampiai, X ir Y lyginami vienos sekundės trukmės kalbos fragmentai ir $C_X^x = \{c_i^x\}_{i=1}^I$ ir $C_Y^y = \{c_i^y\}_{i=1}^I$ stačiakampyje B_i esančių požymių skaičius. Pagal apibrėžimą visi c_i^x ir c_i^y atitinka X ir Y kalbos įrašus ir kadrai priklauso $[x, x+1)$ ir $[y, y+1)$ laiko intervalams. Tegul H_X^x ir H_Y^y yra Šenono entropijos C_X^x ir C_Y^y skaitliukų, t. y.,

$$H_X^x = -\sum_{i=1}^{I} c_i^x/|C_X^x| \log_2(c_i^x/|C_X^x|), \tag{20}$$

$$H_Y^y = -\sum_{i=1}^{I} c_i^y/|C_Y^y| \log_2(c_i^y/|C_Y^y|,) \tag{21}$$

$$|C_X^x| = \sum_{i=1}^{I} c_i^x, \quad |C_Y^y| = \sum_{i=1}^{I} c_i^y. \tag{22}$$

Tegul $C_{X,Y}^{x,y} = \{c_i^x + c_i^y\}_{i=1}^I$ žymi jungtinius C_X^x ir C_Y^y skaitliukus ir

$$H_{X,Y}^{x,y} = -\sum_{i=1}^{I} c_i^{x,y}/|C_{X,Y}^{x,y}| \log_2(c_i^x/|C_{X,Y}^{x,y}|) \tag{23}$$

yra $C_{X,Y}^{x,y}$ jungtinė entropija. Nesunku įrodyti tokį teiginį apie sąryšį tarp šių trijų entropijų.

1 teiginys. *Bet kuriems skaitliukams C_X^x, C_Y^y ir jų jungtiniam $C_{X,Y}^{x,y}$ teisinga tokia nelygybė:*

$$pH_X^x + qH_Y^y \le H_{X,Y}^{x,y} \le pH_X^x + qH_Y^y + H_{p,q}, \tag{24}$$

pur

$$p = \frac{|C_X^x|}{|C_{X,Y}^{x,y}|}, \quad q = \frac{|C_Y^y|}{|C_{X,Y}^{x,y}|} = 1 - p, \tag{25}$$

ir

$$H_{p,q} = -p \log_2 p - q \log_2 q. \tag{26}$$

Proof. Kairioji (24) nelygybės pusė išplaukia iš

$$-\alpha \log_2 \alpha - \beta \log_2 \beta \leq -(\alpha + \beta) \log_2(\alpha + \beta) \quad \alpha > 0, \quad \beta > 0,$$

nelygybės. Dešinioji (24) nelygybės pusė gali būti pagrįsta tokiais informacijos teoprijos argumentai. $H_{X,Y}^{x,y}$ yra Šenono vidutinis informacijos kiekis kurį suteikia atsitiktinai pasiroęs simbolis iš teksto su $C_{X,Y}^{x,y}$ raidžių skaitliukais. Informacija apie pasirodžiusią to paties teksto simbolį galima gauti ir tokiu nebūtinai optimaliu būdu. Pirma klausia *ar pasirodęs simbolis yra iš teksto su* C_X^x *ar* C_Y^y *skaitliukais*? Po to, priklausomai nuo atsakymo į pirmąjį klausimą, su tikimybe p pateikiame antrąjį klausimą *kuris simbolis yra iš teksto su* C_X^x *skaitliukais?* arba su tikimybe $q = 1 - p$ klausiame *kuris simbolis yra iš teksto su* C_Y^y *skaitliukais?*. Atsakymas į pirmąjį klausimą suteikia vidutiniškai $H_{p,q} = -p \log_2 p - q \log_2 q$ bitų informacijoa, o antrasis suteikia H_X^x arba H_Y^y bitų informacijos atitinkamai su tikimybe p ir q. Kadangi bendru atveju pateiktų dviejų klausimų strategija nėra optimali, gauname dešiniąją (24) nelygybę. Griežtą matematinį šios nelygyubės įrodymą paliekame skaitytojui.

1 apibrėžimas. *X įrašo $[x, x+1)$ laiko intervalo (sekundėmis) kalbos fragmento panašumas ρ į Y įrašo $[y, y+1)$ intervalą yra apibrėžiamas formule*

$$\rho(X_{[x,x+1)}, Y_{[y,y+1)}) = 1 + \frac{pH_X^x + qH_Y^y - H_{X,Y}^{x,y}}{H_{p,q}}. \tag{27}$$

Iš 1 išplaukia, kad bet kurių intervalų $X_{[x,x+1)}$ ir $Y_{[y,y+1)}$ panašumas yra visuomet neneigiamas ir neviršija 1. Kitas apibrėžimas skirtas įvertinti $Y_{[y,y+1)}$ kalbos fragmento panašumą į visą X įrašą.

2 apibrėžimas. *$Y_{[y,y+1)}$ kalbos fragmento panašumas į X įrašą yra*

$$\rho(X, Y_{[y,y+1)}) = \frac{\sum_{x=0}^{T_X} \rho(X_{[x,x+1)}, Y_{[y,y+1)})}{T_X}, \tag{28}$$

kur T_X yra X įrašo trukmė sekundėmis.

Kitaip tariant panašumas $\rho(X, Y_{[y,y+1)})$ yra vidutinis $Y_{[y,y+1)}$ fragmento panašumas į aibę visų vienos sekundės trukmė $X_{[x,x+1)}$ intervalų.

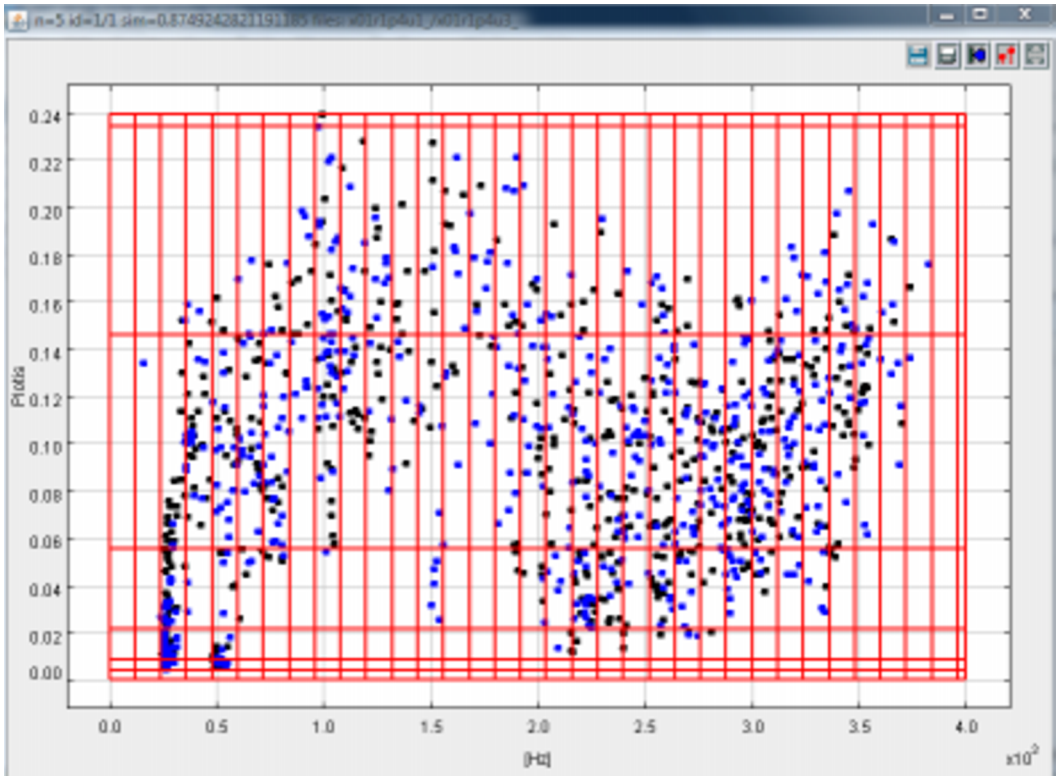
Paskutinis apibrėžimas apibrėžia integruotą X ir Y įrašų panašumą.

3 apibrėžimas. *X ir Y kalbos įrašų panašumas yra skaičius*

$$\rho(X, Y) = \textit{vidutinė reikšmė pusės didžiausiųjų } \rho(X, Y_{[y,y+1)}), y = 0, 1 \ldots, T_Y - 1. \tag{29}$$

Pateikta panašumo metrika $\rho(X, Y)$ yra asimetrinė (bendru atveju $\rho(X, Y) \neq \rho(Y, X)$). Tai motyvuota asimetrija X ir Y duomenyse: X įrašas yra vieno kalbančijojo, o Y įraše gali kalbėti du ar daugiau asmenų. Jei apriori Y įraše kalba taip pat tik vienas asmuo, $\rho(X, Y)$ metriką galima modifikuoti į simetrinę praleidžiant apibrėžime *pusė didžiausiųjų*.

Visos pateiktos kalbos įrašų panašumo metrikos grįstos tarpusavio informacija, yra neneigiamos ir neviršija 1. Jei X ir Y yra absoliučiai skirtingi, t.y. visi X ir Y požymiai patenka į skirtingus B_i stačiakampius, tai visiems x ir y $H^{x,y}_{X,Y} = pH^x_X + qH^y_Y + H_{p,q}$ ir $\rho(X, Y) = 0$. Kitu kraštutiniu atveju, kai visi požymių skaitliukai yra proporcingi ($\forall x, y, i : c^x_i = \text{const } c^y_i$), $H^{x,y}_{X,Y} = pH^x_X + qH^y_Y = H^x_X$ ir $\rho(X, Y) = 1$. Todėl kalbos įrašų panašumo metrika $\rho(X, Y)$ gali būti interpretuojama tikimybiškai: $\rho(X, Y)$ atspindi tikimybę su kuria X lyginamasis dalyvauja Y dialoge.



9 paveikslėlis. x01r1p4u1.wav ir x01r1p4u3.wav vienos sekundės trukmės kadrų požymių skirstiniai.

9 paveikslėlis iliustruoja dviejų kalbos požymių (šviesesni ir tamsesni taškai) skirstinius. Šių skirstinių panašumas yra 0.875.

6 Tikėtinumo santykio logaritmas

Lyginamų porų balso panašumo galimų reikšmių sritis priklauso nuo lyginimo metrikos. Tai apsunkina panašumo reikšmių interpretaciją. Tarkime jei vienos poros balsų panašumo reikšmė yra $\rho = 0.8$, o kitos - $\rho = 0.65$, galime teigti, kad santykinai pirmosios poros balsai labiau panašūs nei antrosios poros balsai, tačiau į kiekybinį klausimą "kiek kartų labiau panašūs" atsakyti negalima. Todėl kriminalistikoje įsivyravo *tikėtinumo santykio* (angl. *Likelihood Ratio (LR)*) metrika, kurios reikšmes galima interpretuoti kiekybiškai. Balsų poros (X, Y) panašumo tikėtinumo santykis apibrėžiamas formule

$$LR(X,Y) = \frac{\text{Tikėtinumas, kad lyginami poros } (X,Y) \text{ balsai sutampa}}{\text{Tikėtinums, kad lyginami poros } (X,Y) \text{ balsai nesutampa}}. \quad (30)$$

Bendrąja prasme tikėtinumas yra modelio tikimybė, kai žinomi atlikto eksperimento rezultatai. Mūsų atveju "eksperimento rezultatai" yra balsai X ir Y ir jų požymiai. Skaitiklio tikėtinumas yra kokio nors modelio tikimybė gauti ("išmatuoti") X ir Y balsų požymius, darant prielaidą, kad balsai yra to paties asmens, o vardiklio tikėtinumas yra tikimybė gauti tuos pačius X ir Y požymius, darant prielaidą, kad balsai yra skirtingų asmenų. Pagal apibrėžimą LR reikšmės gali kisti nuo 0 iki ∞. Jei $0 \leq LR = LR(X,Y) < 1$, labiau tikėtina, kad lyginami balsai X ir Y yra skirtingi. Priešingai, jei $1 < LR < \infty$, labiau tikėtina, kad lyginami balsai sutampa. Kad išvengti intervalų $(0, 1)$ ir $(1, \infty)$ ilgių asimetrijos, dažnai vartotojui pateikiama LR natūraliojo logaritmo reikšmė, kuri sutrumpintai žymima $LLR(X,Y) = LLR = \log(LR(X,Y))$ *(angl. Log Likelihood Ratio, LLR*). Jei tikėtinumo santykio logaritmas teigiamas, labiau tikėtina, kad lyginamos poros balso pavyzdžiai yra vieno asmens. Priešingu atveju, kai tikėtinumo santykio logaritmas yra neigiamas, labiau tikėtina, kad tiriamos poros balso pavyzdžiai priklauso skirtingiems asmenims.

1 lentelėje pateiktos LR ir LLR reikšmės bei jų interpretacija. Pavyzdžiui, jei LLR = 6.9, tai apie 1000 kartų labiau tikėtina, kad lyginamojo ir tiriamojo balso pavyzdžiai X ir Y priklauso vienam ir tam pačiam asmeniui nei skirtingiems ir atvirkščiai, jei LLR = -6.9, tai apie 1000 kartų labiau tikėtina, kad lyginamojo ir tiriamojo balso pavyzdžiai priklauso skirtingiems asmenims nei tam pačiam asmeniui. Tikėtinumo santykiui įvertinti naudojami visi turimi lyginamieji balso pavyzdžiai ir vieno tiriamųjų katalogo balso pavyzdžiai. Todėl kuo lyginamųjų daugiau, tuo LLR patikimumas didesnis.

Kad pagal (30) apskaičiuoti tikėtinumo santykį LR, reikia įvertinti skaitiklio ir vardiklio reikšmes. Tikėtinumo reikšmės priklauso nuo pasirinkto modelio. Literatūroje modeliai dažnai konstruojami remiantis Gauso skirstiniu. Gauso skirstinys yra simetrinis ir su bet kokiais modelio parametrais modeliuojamas dydis gali įgyti bet kokias reikšmes iš intervalo $(-\infty, \infty)$. Tačiau mūsų atveju modeliuojamos panašumo reikšmės ρ visuomet patenka į $[0, 1]$ uždarą intervalą. Todėl mes modelio pagrindu pasirinkome eksponentinį skirstinį. Laikome, kad kiekvieno fiksuoto tiriamojo X panašumo reikšmių tikėtinumai tenkina tokias

1 lentelė. Tikėtinumo santykio LR ir jo natūraliojo logaritmo LLR reikšmės

LR	LLR	Interpretacija
1000	6.9	Labai tikėtina, kad lyginami balsai sutampa
403.4	6	Labai tikėtina, kad lyginami balsai sutampa
100	4.6	Pakankamai tikėtina, kad lyginami balsai sutampa
20.1	3	Tikėtina, kad lyginami balsai sutampa
10	2.3	Labiau tikėtina, kad lyginami balsai sutampa
7.4	2	Labiau tikėtina, kad lyginami balsai sutampa
1	0	Vienodai tikėtina, kad lyginami balsai sutampa arba nesutampa
1/7.4	-2	Labiau tikėtina, kad lyginami balsai nesutampa
1/10	2.-3	Labiau tikėtina, kad lyginami balsai nesutampa
1/20.1	-3	Tikėtina, kad lyginami balsai nesutampa
1/100	-4.6	Pakankamai tikėtina, kad lyginami balsai nesutampa
1/403.4	-6	Labai tikėtina, kad lyginami balsai nesutampa
1/1000	-6.9	Labai tikėtina, kad lyginami balsai nesutampa

lygtis:

$$P(\rho(X,Y)=s|\text{esant prielaidai, kad } X \text{ ir } Y \text{ yra vienodi}) = \lambda \exp(\lambda(s-a_X)) \tag{31}$$

ir

$$P(\rho(X,Y)=s|\text{esant prielaidai, kad } X \text{ ir } Y \text{ yra skirtingi}) = \lambda \exp(-\lambda(s-b_X)) \tag{32}$$

Laikome, kad ir skaitiklio ir vardiklio modelio parametras λ yra vienodas. Šį parametrą galima įvertinti remiantis panašumo reikšmių dispersija, nes $\frac{1}{\lambda^2}$ yra eksponentinio skirstinio dispersija. Poslinkio parametrai $0 < a_X < b_X < 1$ parenkami individualiai kiekvienam tiriamajam X. Jei šie paramtrai yra žinomi ir poros (X,Y) apskaičiuota panašumo reikšmė $\rho(X,Y)=s$, tuomet gauname tokias paprastas formules:

$$LR(X,Y) = \frac{\lambda \exp(\lambda(s-a_X))}{\lambda \exp(-\lambda(s-b_X)} = \exp(\lambda(2s-a_X-b_X)) \tag{33}$$

ir

$$LLR(X,Y) = 2\lambda(s-\frac{a_X+b_X}{2}). \tag{34}$$

Aprašysime modelio parametrų λ, a_X ir b_X įvertinimo procedūrą.

6.1 λ parametro įvertis

$\frac{1}{\lambda^2}$ parametro tikimybinė interpretacija yra balsų panašumo reikšmių dispersija skaičiuojant ją atskirai vienodiems ir skirtingiems balsams. Kadangi lyginamiems balsas turima tiksli informacija kokie balsai sutampa, o kokie skirtingi, λ parametras įvertinamas naudojant tik lyginamųjų balsus $Y_1, Y_2, \cdots, Y_L$. Žymėjimų

formulėse paprastumo dėlei laikysime, kad visi lyginamieji balsai $Y_1, Y_2, \cdots, Y_L$ yra skirtingų asmenų. Tuomet apskaičiuojame visus galimus $L^2 - L$ skirtingų balsų panašumus:

$$\rho(Y_i, Y_j) = s_{i,j}, \quad i, j = 1, \cdots, L, i \neq j.$$

Toliau kiekvienoje eilutėje i išsirenkame $K = [\sqrt(L)]$ didžiausiųjų:

$$S_{i,k}, \quad k = 1, 2, \cdots, K$$

ir dispersiją $\frac{1}{\lambda^2}$ įvertiname pagal įprastą dispersijos įverčio formulę:

$$\frac{1}{\lambda^2} = \sum_{i=1}^{L}(\sum_{k=1}^{K} S_{i,k}^2 - (\sum_{k=1}^{K} S_{i,k})^2/K)/(K-1))/L.$$

6.2 a_X ir b_X parametrų įvertis

a_X ir b_X ir parametrų įvertis priklauso nuo tiriamojo X ir nuo to ar lyginami balsai taria *vienodą* ar *skirtingą* teksą (angl. *text dependent and text independent*). Tarkime lyginami balsai taria tą patį tekstą. Fiksuojame tiriamąjį balsą X ir apskaičiuojame jo panašumą į visus lyginamuosius $Y_1, Y_2 \cdots Y_L$:

$$s_1, s_2, \cdots, s_L \quad (s_l = \rho(X, Y_l)).$$

Laikome kad didžiausias panašumas, t.y. $s^{max} = \max_l s_l$, priklauso vieno asmens lygintų balsų porai ir jį naudojame skaitiklio tikėtinumo modelio parametrui a_X įvertinti, postuluojant, kad $a_X = s^{max}$. Laikome, kad antroji pagal dydį balsų poros panašumo reikšmė $s^{sec} = \max_l\{s_l$ išskyrus $s^{max}\}$ atitinka skirtingus asmenis ir ją naudojame apibrėžiant tikėtinumo santykio vardiklio eksponentinio skirstinio modelio parametrą b_X, postuluojant, kad $b_X = s^{sec}$.

Jei tariami tekstai yra skirtingi, aprašyta procedūra yra modifikuojama įvertinant a_X dviejų dižiausiųjų panašumų aritmetiniu vidurkiu, o b_X prilyginamas trečiajam pagal dydį panašumui.

7 Sistemos tyrimas

7.1 Tyrimo duomenys ir rezultatai

Kad palyginti pasiūlytą diktoriaus identifikavimo techniką su kitomis naudojome rusų kalbos duomenų bazę (RUSBASE), kurią pateikė Lietuvos Teismo Ekspertizės Fonoskopinių Ekspertizių skyrius. Duomenų specifikacija pateikta ELRA (European Language Resources Association) [7] šaltinyje. Tyrimams taip pat buvo panaudota nepriklausoma nuo kalbančiojo Netherlands Forensic Institute Speaker Recognition Evaluation (NFISRE) duomenų bazė. NFISRE 2004-2005 atliko tyrimą vertinant įvairius Europos Sąjungos fonoskopinių teismo ekspertizės centrų naudojamus algoritmus. NFISRE yra du lyginamojo įrašai. Tiriamųjų

įrašai yra nuo 20 sek. iki 10 min. trukmės dialogai. NFISRE užduoti nustatyti ar pateiktuose tiriamuose įrašuose dalyvauja lyginamasis. Kad išgryninti lyginamąjį, 2-uose pateiktuose lyginamojo įrašuose rankiniu būdu ištrynėme pokalbyje dalyvavusio kito asmens kalbos fragmentus. Tiriamieji įrašai buvo tikrinami pilnai automatiškai naudojant sukurtą balso identifikavimo sistemą. Lyginant su pateikais NFI [2] atsakymais apie tiriamuosius mums pavyko gauti idealų atpažinimą, t.y. atsirado panašumo slenkstis kuris pilnai atskyrė visus tiriamųjų įrašus, kuriuose kalbėjo lyginamasis nuo likusių tiriamųjų, kuriuose lyginamasis nekalbėjo.

RUSBASE yra 5-ių skirtingų sakinių įrašai su vidutiniškai 15 sesijų vienam sakiniui, bazėje yra 44 vyrų 35 moterų balsai, bendras kalbos įrašų kiekis apie 500 Mb. Pirmosios trys sesijos naudojamos mokymui (lyginimui), likusios testavimui (tyrimui).

2 lentelė. RUSBASE atpažinimo rezultatai, 1-as sakinys, vyrų balsas, naudojamos skirtingos balsų palyginimo technikos

Metodas	Savybės	EER [%]
VQ	MFCC	8.8
GMM	MFCC	5.8
GMM	F&A	5.1
Foneminis	F&A	2.32

RUSBASE perdavimo funkcijos faze grįsti klasifikavimo rezultatai buvo palyginti su Gauso Mišinių Modeliu (GMM) naudojančiu Mel skalės kepstro koeficientus(MFCC), formantes ir antiformantes (F&A), pagrindinio tono reikšmę F0. 2 lentelėje pateikti klasifikavimo lygios klaidos (EER) reikšmės RUSBASE atveju, pirmas sakinys, vyrų balsass, naudojant MFCC, F&A, ir F0 požymius ir vektorinio kvantavimo (VQ) bei GMM klasifikavimo technikas. EER kinta nuo 2.32 iki 8.8% (žiūr. 2 lentelę). Mūsų grupinės delsos ekstremumų požymiais ir tarpusavio informacija paremta balso identifikavimo sistema tiems patiems duomenims turėjo EER=$0,042\%$. 3 lentelėje pateikti mūsų sistemos rezultatai likusiems RUSBASE duomenims. Čia FAR0 ir FRR0 atitinkamai mulinės priėmimo klaidos (Zero False Acceptance Rate) nulinės atmetimo klaidos (Zero False Rejection Rate) reikšmės.

Literatūra

[1] Cai Jinhai, Jian Gangji and Zhang Lihe,: New method for extracting speech formants using LPC phase spectrum. Electronic letters, Vol.29, Nr 24, 2081-2082, (1993)

3 lentelė. Balso atpažinimas naudojant grupinės delsos požymius ir tarpusavio informacija grįstą panašumą. RUSBASE duomenų bazė, 1–5 sakiniai

Sakinys	Balsas	FAR0 [%]	EER0 [%]	FRR0
1	man	1.8	0.042	0.12
1	woman	1.96	0.042	0.07
2	man	0.8	0.084	0.12
2	woman	2.17	0.2	1.37
3	man	3.19	0.058	0.09
3	woman	1.96	0.033	0.06
4	man	0.6	0.01	0.02
4	woman	4.6	0.112	0.15
5	man	2.79	0.199	0.59
5	woman	0.44	0.007	0.01

[2] Gambier-Langeveld, T. : Netherlands Forensic Institute [NFD, speaker recognition fake case evaluation. June 2-3, 2005, 8th Meeting of ENFSI Expert Working Group for Forensic Speech and Audio Analysis.

[3] An-Tze Yu, Hsiao-Chuan Wang, : Channel Effect Compensation in LSF Domain, EURASIP Journal on Applied Signal Processing, 9, 922-929 (2003)

[4] D. A. Reynold, R. C. Rose, : Robust Text-Independent Speaker Identification Using Gaussiam Mixture Speakers Models, IEEE transactions on speech and audio processing, 3:1, January, 72-83, (1995)

[5] F. Itakura and S. Saito,, : Analysis synthesis telephony based upon the maximum likelihoood method. Reports on 6th Int. Cong. Acoust., ed. By Y Kohasi, Tokyo, C-5-5, C17-20 (1968)

[6] H. W. Strube,: Linear prediction on a warped frequency scale, J. Acoust. Soc. Am., 68 : 4, 1071==1076, October (1980)

[7] http://www.linguistlist.org/issues/9/9-891.html, ELRA-S0050 Russian speech database (STC)