VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Juozas KAMARAUSKAS

ASMENS ATPAŽINIMAS PAGAL BALSĄ

DAKTARO DISERTACIJA

TECHNOLOGIJOS MOKSLAI,
INFORMATIKOS INŽINERIJA (07T)

VGTU
LEIDYKLA
Vilnius TECHNIKA 2009

Disertacija rengta 2004–2009 metais Matematikos ir informatikos institute.

Mokslinis vadovas

doc. dr. Antanas Leonas LIPEIKA (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T).

http://leidykla.vgtu.lt
VGTU leidyklos TECHNIKA 1610-M mokslo literatūros knyga

ISBN 978-9955-28-422-2

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Juozas KAMARAUSKAS

SPEAKER RECOGNITION BY VOICE

DOCTORAL DISSERTATION

TECHNOLOGICAL SCIENCES,
INFORMATICS ENGINEERING (07T)

VGTU
LEIDYKLA
Vilnius TECHNIKA 2009

Doctoral dissertation was prepared at Institute of Mathematics and Informatics in 2004–2009.

Scientific Supervisor

Assoc Prof Dr Antanas Leonas LIPEIKA (Institute of Mathematics and Informatics, Technological Sciences, Informatics Engineering – 07T).

Santrauka

Disertacijoje nagrinėjami kalbančiojo atpažinimo pagal balsą klausimai. Aptartos kalbančiojo atpažinimo sistemos, jų raida, atpažinimo problemos, požymių sistemos įvairovė bei kalbančiojo modeliavimo ir požymių palyginimo metodai, naudojami nuo ištarto teksto nepriklausomame bei priklausomame kalbančiojo atpažinime.

Darbo metu sukurta nuo ištarto teksto nepriklausanti kalbančiojo atpažinimo sistema. Kalbėtojų modelių kūrimui ir požymių palyginimui buvo panaudoti Gauso mišinių modeliai.

Pasiūlytas automatinis vokalizuotų garsų išrinkimo (segmentavimo) metodas. Šis metodas yra greitai veikiantis ir nereikalaujantis iš vartotojo jokių papildomų veiksmų, tokių kaip kalbos signalo ir triukšmo pavyzdžių nurodymas.

Pasiūlyta požymių vektorių sistema, susidedanti iš žadinimo signalo bei balso trakto parametrų. Kaip žadinimo signalo parametras, panaudotas žadinimo signalo pagrindinis dažnis, kaip balso trakto parametrai, panaudotos keturios formantės bei trys antiformantės. Siekiant suvienodinti žemesnių bei aukštesnių formančių ir antiformančių dispersijas, jas pasiūlėme skaičiuoti melų skalėje. Rezultatų palyginimui sistemoje buvo realizuoti standartiniai požymiai, naudojami kalbos bei asmens atpažinime – melų skalės kepstro koeficientai (MSKK). Atlikti kalbančiojo atpažinimo eksperimentai parodė, kad panaudojus pasiūlytą požymių sistemą buvo gauti geresni atpažinimo rezultatai, nei panaudojus standartinius požymius (MSKK). Gautas lygių klaidų lygis, panaudojant pasiūlytą požymių sistemą, – 5,17 %, tuo tarpu panaudojant MSKK – 5,86 %. Formančių skaičiavimas melų skalėje taip pat šiek tiek pagerino atpažinimo rezultatus, nei jų skaičiavimas tiesinėje skalėje.

Pasiūlyta požymių sistema yra mažesnės dimensijos ir susideda iš 8 komponenčių, tuo tarpu standartiniai požymiai – MSKK susideda iš 13 komponenčių. Dėl šių priežasčių, kuriant kalbėtojų modelius bei atpažinimo metu naudojant pasiūlytą požymių sistemą, reikia atlikti maždaug 1,6 karto mažiau skaičiavimo operacijų.

Formančių bei antiformančių įvertinimui panaudotas spektrinių porų metodas, kadangi ne visada galima jas tiesiogiai rasti.

Pasiūlytas metodas pradiniam GMM parametrų vertinimui. Pradiniai GMM parametrai apskaičiuojami padalinus pradinę požymių vektorių aibę į klasterius bei radus atitinkamų klasterių statistinius parametrus. Dėl to klasterių formavimui pasiūlėme naudoti vektorinio kvantavimo algoritmą. Sistemoje realizuoti ir kiti klasterių formavimo metodai: tiesinis požymių vektorių dalijimas į klasterius bei atsitiktinis klasterių formavimas. Atlikus eksperimentus paaiškėjo, kad panaudojant vektorinio kvantavimo metodą buvo gauti geriausi atpažinimo rezultatai, lygių klaidų lygis sumažėjo 0,71 %, lyginant su atsitiktinio klasterių formavimo metodu, bei 0,88 %, lyginant su tiesinio dalijimo į klasterius metodu, tačiau nesumažino iteracijų skaičiaus, reikalingo tikslinant kalbėtojų modelius.

Abstract

Questions of speaker's recognition by voice are investigated in this dissertation. Speaker recognition systems, their evolution, problems of recognition, systems of features, questions of speaker modeling and matching used in text-independent and text-dependent speaker recognition are considered too.

The text-independent speaker recognition system has been developed during this work. The Gaussian mixture model approach was used for speaker modeling and pattern matching.

The automatic method for voice activity detection was proposed. This method is fast and does not require any additional actions from the user, such as indicating patterns of the speech signal and noise.

The system of the features was proposed. This system consists of parameters of excitation source (glottal) and parameters of the vocal tract. The fundamental frequency was taken as an excitation source parameter and four formants with three antiformants were taken as parameters of the vocal tract. In order to equate dispersions of the formants and antiformants we propose to use them in mel-frequency scale. The standard mel-frequency cepstral coefficients (MFCC) for comparison of the results were implemented in the recognition system too. These features make baseline in speech and speaker recognition. The experiments of speaker recognition have shown that our proposed system of features outperformed standard mel-frequency cepstral coefficients. The equal error rate (EER) was equal to 5.17% using proposed features system compared to 5.86% that has been obtained using standard MFCC. Usage of the formants and antiformants in the mel-frequency scale improved recognition accuracy in comparison to usage of those in linear scale.

The dimension of proposed system of features is lower and these features consist of 8 components, meanwhile standard features (MFCC) consist of 13 components. Therefore we need to implement 1.6 times less operations of calculation when we create speaker's models or during the recognition, using proposed system of features compared to standard MFCC.

The method of line spectral pairs was used for approximate calculation of formants and antiformants, because they are not always easy to be found directly.

The method of estimation of initial GMM parameters was proposed too. Initial parameters of GMM are calculated after division of the initial space of the feature vectors into clusters. Statistical parameters of the clusters are calculated and assigned to corresponding Gaussian mixture as initial parameters. Vector quantization approach was proposed for this case. Other methods of forming of clusters are implemented in this system too: linear division of feature vectors into the clusters and random forming of the clusters. Experiments performed have shown that vector quantization approach provided best results of accuracy in this case and outperformed other methods of forming of clusters. Method of random forming was outperformed by 0.71% and method of linear division – by 0.88%, yet not reducing count of iterations necessary to build speaker's model.

Žymėjimai

Simboliai

a_i – *i*-tasis tiesinės prognozės modelio koeficientas,
C – kodinė knyga,
$c(k)$ – *k*-tasis kepstro koeficientas,
$d(r,z)$ – atstumas tarp vektorių *r* ir *z*,
$F0$ – žadinimo signalo pagrindinis dažnis,
F_d – diskretizacijos dažnis,
G – tiesinės prognozės modelio stiprinimo koeficientas,
$H(i,k)$ – trikampių filtrų funkcija,
K – signalo kadrų skaičius,
M – Gauso mišinių komponenčių skaičius,
N – signalo kadro ilgis atskaitomis,
p_i – *i*-tosios Gauso mišinio komponentės svorio koeficientas,
$s(j,n)$ – *j*-tojo kadro diskretinio laiko signalas,
$s(n)$ – diskretinio laiko signalas,
$w(n)$ – diskretinio laiko lango funkcija,
X – požymių vektorių seka,
$X(k)$ – diskrečioji Furjė transformacija,
x_i – *i*-tasis požymių vektorius.

Santrumpos

ADFT – atvirkštinė diskrečioji Furjė transformacija,
ALA – Apibendrintas Lloydo algoritmas (*angl. GLA – Generalized Lloyd algorithm*),
AVM – atraminių vektorių mašinos,
DACh – dažninė amplitudės charakteristika,
DKT – diskrečioji kosinusų transformacija,
DET – *angl. – Detect Error trade-off*,
DFT – diskrečioji Furjė transformacija,
DNT – dirbtinių neuronų tinklai,
DLK – dinaminis laiko skalės kraipymas,
KP – klaidingas priėmimas (*angl. FA – False accept*),
KPL – klaidingo priėmimo lygis (*angl. FAR – False accept rate*),
KA – klaidingas atmetimas (*angl. FR – False reject*),

KAL – klaidingo atmetimo lygis (*angl. FRR – False reject rate*),
GFT – greitoji Furjė transformacija,
GMM – Gauso mišinių modeliai,
LBG – Linde, Buzo, Gray,
LKL – lygių klaidų lygis arba tikimybė,
MSKK – melų skalės kepstro koeficientai (*angl. MFCC – Mel-frequency cepstrum coefficients*),
MSSK – melų skalės spektro koeficientai (*angl. MFSC – Mel-frequency spectrum coefficients*),
MVM – matematinės vilties maksimizavimas,
PMM – paslėptieji Markovo modeliai,
RIR – ribota impulsinė reakcija,
TPM – tiesinės prognozės modelis,
TPMK – tiesinės prognozės modelio kepstras,
VK – vektorinis kvantavimas.

Turinys

Contents

Įvadas

Tiriamoji problema

Šiame darbe nagrinėjamos kalbančiojo asmens atpažinimo pagal balsą problemos, naudojamos požymių sistemos bei jų palyginimo metodai, automatinio kalbos signalų segmentavimo klausimai.

Darbo aktualumas

Šiuolaikiniame pasaulyje vis aktualesnės tampa asmens atpažinimo pagal balsą problemos. Šios problemos atsiranda kriminalistikoje (kai reikia identifikuoti kalbantįjį asmenį, pvz. turint telefoninį nusikaltėlio pokalbį arba kriminalinėje paieškoje), informacijos apsaugoje (pvz. rinkmenų užkodavime), tai gali būti taikoma įėjimo kontrolės sistemose, internetinėje prekyboje ir t. t. Joms yra skiriamas didelis dėmesys, materialiniai bei intelektualiniai ištekliai, sukurti įvairūs testavimo centrai. Jei kitos biometrijos rūšys reikalauja specialios, dažnai brangiai kainuojančios įrangos (tarkime akies rainelės ar pirštų antspaudų skaitytuvas), asmens atpažinimo pagal balsą sistemos to nereikalauja. Dėl šių priežasčių asmens identifikavimo pagal balsą algoritmų kūrimui visame pasaulyje skiriamas labai didelis dėmesys ir, pagal prognozes, laukiama įvairių

balso biometrijos problemų sprendimų kriminalistikoje, mobiliojoje bankininkystėje bei internetinėje prekyboje.

Nepaisant nemažų pasiekimų šioje srityje, iki šiol nėra sukurtos nei teorijos, kaip žmogus atskiria vieną balsą nuo kito akustiniame lygyje, nei universalios požymių sistemos, leidžiančios laisvai atskirti skirtingus balsus, esant skirtingoms frazėms, skirtingai kalbėjimo aplinkai, skirtingiems garso įrašymo kanalams, triukšmams ir t. t. Asmens atpažinimo pagal balsą sistemos gana gerai veikia tada, kai yra naudojamos tos pačios frazės, kontroliuojamos įrašymo sąlygos, didelis santykis signalas – triukšmas.

Visame pasaulyje, o ypač Lietuvoje, didesnis dėmesys yra skiriamas kalbos atpažinimo sistemų kūrimui. Reiktų paminėti, kad šiuo metu kalbančiojo atpažinime plačiausiai naudojamos požymių sistemos yra tos pačios kaip ir kalbos atpažinime, tačiau tai yra du skirtingi uždaviniai.

Taip pat reiktų paminėti, kad iš visų biometrijos rūšių (asmens atpažinimo pagal jo anatomines bei fiziologines savybes), naudojant balso biometriją kol kas gaunami vieni iš prasčiausių rezultatų, tačiau ateityje ši biometrijos rūšis galėtų turėti labai platų pritaikymą. Dėl šios priežasties reikėtų atlikti daugiau tyrimų šioje srityje, ieškoti naujų kalbos signalo požymių, leidžiančių vienareikšmiškai nustatyti asmenį, taip pat spręsti įrašymo sąlygų neatitikimo, triukšmų ir įrašymo kanalo įtakos sumažinimo problemas, kurios tiesiogiai yra susijusios su kalbos signalo kokybe.

Tyrimų objektas

Darbo tyrimų objektas – asmens atpažinimas pagal balsą ir kalbos signalų apdorojimas.

Darbo tikslas

Pagrindinis šio darbo tikslas – atlikti kalbančiojo atpažinimo sistemų analizę, pasiūlyti sprendimus, didinančius kalbančiojo atpažinimo sistemos veikimo tikslumą bei darbo efektyvumą.

Darbo uždaviniai

Darbo tikslui pasiekti darbe reikia spręsti šiuos uždavinius:

1. Pasiūlyti automatinį vokalizuotų garsų išskyrimo iš įrašyto kalbos signalo algoritmą.
2. Pasiūlyti naują efektyvią požymių sistemą, didinančią asmens atpažinimo tikslumą bei mažinančią reikalingų skaičiavimo operacijų skaičių.
3. Pasiūlyti efektyvų metodą kalbėtojų modelių pradinių parametrų vertinimui.
4. Realizuoti pasiūlytus metodus. Eksperimentiškai įvertinti sukurtos atpažinimo sistemos tikslumą, lyginant pasiūlytus požymius su šiuo metu vienais iš plačiausiai naudojamų pasaulyje.

Tyrimų metodika

Teorinei analizei panaudotos matematikos, taip pat tikimybių teorijos bei matematinės statistikos, skaitmeninio signalų apdorojimo bei atpažinimo teorijos žinios.

Darbo mokslinis naujumas ir jo reikšmė

Rengiant disertaciją buvo gauti šie informatikos inžinerijos mokslui nauji rezultatai:

1. Sukurtas automatinis vokalizuotų garsų išrinkimo iš kalbos bei triukšmo signalų metodas, veikiantis tiksliau nei, pavyzdžiui, energijos slenksčio metodas. Pasiūlytas metodas yra kompleksinis, susidedantis iš kelių atskirų algoritmų: signalo kadrų su nulinėmis ir labai žemomis signalo reikšmėmis atmetimo, foninio triukšmo radimo bei melų skalės spektro slenksčio nustatymo, žadinimo signalo radimo bei nevokalizuotų garsų atmetimo. Šis metodas taip pat pašalina įvairius pavienius impulsinius trikdžius. Kalbančiųjų modelių kūrimui bei atpažinimui parenkami tik vokalizuoti garsai.
2. Pasiūlyta nauja požymių vektorių, turinčių nedaug komponenčių, sistema. Kaip žinoma, kalbos signalas generuojamas žadinimo signalui veikiant balso traktą. Pasiūlyti požymių vektoriai susideda tiek iš žadinimo signalo parametrų, tiek ir iš balso trakto parametrų. Kaip žadinimo signalo parametras yra naudojamas žadinimo signalo pagrindinis dažnis (F0), kaip balso trakto parametrai – formantės (kalbos signalo kadro Furjė spektro gaubtinės maksimumų dažniai) bei antiformantės

(kalbos signalo kadro Furjė spektro gaubtinės minimumų dažniai). Siekiant sumažinti aukštesnių formančių bei antiformančių dispersiją, jos skaičiuojamos melų skalėje. Kadangi pasiūlytų požymių vektorių komponenčių skaičius yra nedidelis (nuo penkių iki aštuonių vektoriaus komponenčių), lyginant su tradicinėmis (trylika arba trisdešimt devynios komponentės), dėl to gerokai pagreitėja skaičiavimai, ypač pakartotiniame parametrų vertinime, matematinės vilties maksimizavimo algoritme, kuriant kalbėtojų Gauso mišinių modelius.

3. Pasiūlytas metodas pradiniam kalbančiųjų modelių parametrų vertinimui. Tam tikslui panaudotas modifikuotas LBG vektorinio kvantavimo algoritmas.
4. Naudojant *TURBO C++ 2006* integruotą programų kūrimo aplinką C++ kalba sukurta programinė įranga, leidžianti atlikti kalbančiojo atpažinimo tyrimus bei vykdyti asmenų paiešką balsų bazėse.

Darbo rezultatų praktinė reikšmė

Tyrimų rezultatai gali būti naudojami įvairių automatinio kalbančiojo atpažinimo sistemų projektavimui. Pradinių GMM parametrų vertinimo metodas gali būti panaudotas kitų klasifikatorių, naudojančių Gauso mišinių modelius, pradinių parametrų vertinimui.

Ginamieji teiginiai

1. Pasiūlytoji požymių sistema, susidedanti iš žadinimo signalo bei balso trakto parametrų.
2. Pasiūlytas vokalizuotų garsų išskyrimo metodas.
3. Pasiūlytas pradinių GMM parametrų vertinimo metodas.
4. Sukurtoji automatinio kalbančiojo atpažinimo programinė įranga.

Darbo rezultatų aprobavimas

Disertacijos tema yra atspausdinti 4 moksliniai straipsniai: vienas – mokslo žurnale, įtrauktame į Thomson ISI sąrašą (Kamarauskas 2006); vienas – mokslo žurnale, įtrauktame į WOS ISI sąrašą (Kamarauskas 2008), du – kitose

tarptautinių ir respublikinių konferencijų medžiagose (Kamarauskas 2007; Šalna, Kamarauskas 2005).

Disertacijoje atliktų tyrimų rezultatai buvo paskelbti septyniose mokslinėse konferencijose Lietuvoje ir užsienyje:

- Tarptautinėje konferencijoje „*Elektronika*“, 2005, 2006, 2007, 2008 m., Vilniuje;
- Tarptautinėje mokslinėje – praktinėje konferencijoje „*Kriminalistika ir teismo ekspertizė: mokslas, studijos, praktika*“, 2005 m., Vilniuje;
- Konferencijoje „*Informacinės technologijos 2007*“, 2007 m., Kaune;
- Tarptautinėje konferencijoje „*Bio-Inspired Signal and Image Processing BISIP'08*“, 2008 m., Varšuvoje.

Disertacijos struktūra

Disertaciją sudaro įvadas, keturi skyriai, išvados, literatūros sąrašas ir autoriaus publikacijų sąrašas. Disertacijos aiškinamąjį raštą sudaro 124 teksto puslapiai, su 58 paveikslais ir 8 lentelėmis. Literatūros sąraše 120 šaltinių.

Įvade suformuluojama tiriamoji problema, aptariamas temos aktualumas, darbo tikslas, metodai ir priemonės, mokslinis naujumas, ginamieji teiginiai.

Pirmame skyriuje bendrai aptariamos kalbančiojo atpažinimo sistemos, jų klasifikacija, taip pat pagrindinės sąvokos. Čia taip pat aptariama kalbančiojo atpažinimo sistemų raida, biometrinių sistemų darbingumo vertinimo parametrai, automatinės kalbančiojo atpažinimo sistemos, o taip pat ir pagrindinės problemos, su kuriomis susiduriama kalbančiojo atpažinime.

Antrame skyriuje nagrinėjami kalbos generavimo bei modeliavimo klausimai, detaliai nagrinėjami automatinių kalbančiojo atpažinimo sistemų elementai, kalbos signalų apdorojimo klausimai, požymių sistemos, naudojamos kalbos ir kalbančiojo atpažinime, kalbos signalų segmentavimo klausimai ir kalbėtojų modelių kūrimo bei požymių palyginimo būdai, naudojami nepriklausomame nuo ištartos frazės kalbančiojo atpažinime.

Trečiasis skyrius skirtas kalbančiojo atpažinimo sistemos realizacijai. Čia aptariamas pasiūlytas automatinis vokalizuotų garsų išrinkimo iš kalbos signalų bei triukšmo metodas, šiek tiek modifikuotas žadinimo signalo pagrindinio dažnio radimo metodas, pasiūlyta požymių vektorių sistema bei pradinio GMM parametrų vertinimo metodas.

Ketvirtajame skyriuje pateikti sukurtos atpažinimo sistemos eksperimentinio tyrimo rezultatai. Eksperimentais tirtas kalbančiojo asmens atpažinimo tikslumas, panaudojant įvairias požymių sistemas: pasiūlytą požymių vektorių sistemą, susidedančią iš formančių, antiformančių ir žadinimo signalo pagrindinio dažnio, taip pat ir atskiras šios požymių sistemos dalis. Rezultatų palyginimui atlikti

eksperimentai ir su standartiniais požymiais – melų skalės kepstro koeficientais. Eksperimentų metu tirta ir atpažinimo tikslumo priklausomybė nuo Gauso mišinių komponenčių skaičiaus, pradinio GMM parametrų vertinimo įtaka atpažinimo tikslumui ir t. t.

Paskutiniame skyriuje apibendrinami darbo rezultatai ir suformuluojamos išvados, aptariamos tolesnės atpažinimo sistemos vystymo galimybės.

1

Kalbančiojo atpažinimo sistemos

Šiuo metu yra sukurta daug kalbos technologijų taikymų. Kalbos atpažinimo technologijos gali būti padalintos į tris pagrindines dalis: kalbos atpažinimas, kalbančiojo atpažinimas bei kitas atpažinimas (lyties, nuotaikos, girtumo, amžiaus ir t. t.).

Su asmens atpažinimu mes susiduriame kiekvieną dieną, kalbėdami telefonu, klausydami masinių žiniasklaidos priemonių (televizijos, radijo). Pagal balsą mes galime ne tik atpažinti asmenį, bet ir nustatyti kitas jo savybes: amžių, emocinę būseną, lytį ir t. t. Poreikis identifikuoti asmenį pasaulyje nuolat auga įvairiose srityse. Yra trys pagrindiniai būdai, kaip identifikuoti asmenį (Prabhakar *et al.* 2003):

- Pagal tai, ką asmuo turi (pvz. kortelė, raktas ir pan.).
- Pagal tai, ką asmuo žino (pvz. vartotojo vardas, slaptažodis, PIN kodas ir pan.).
- Pagal asmens fiziologines savybes (akies rainelė, balsas, pirštų antspaudai, veido bruožai, DNR ir t. t.).

Pirmieji du būdai, taip pat dalis iš jų naudojamų jau daug šimtmečių, yra bene labiausiai paplitę, tačiau jie turi ir atitinkamų trūkumų. Kortelė, raktai gali būti pavogti ar pamesti, slaptažodžiai, PIN kodai supainioti ar atspėti kitų asmenų. Paskutinioji autentifikacijos metodų klasė vadinama biometriniu asmens autentifikavimu (Prabhakar *et al.* 2003), kur didelė dalis anksčiau paminėtų

problemų dingsta. Kiekvienas asmuo turi savo unikalią anatomiją, fiziologiją, savo įpročius ir pagal tai jį kiekvieną dieną identifikuoja kiti žmonės.

Nuolat tobulėjant ir pingant kompiuterinei įrangai bei kitai elektroninei technikai, biometrinės technologijos vis plačiau pradedamos taikyti įvairiose gyvenimo srityse.

Balsas skiriasi nuo kitų žmogaus biometrinių savybių, kaip pvz. pirštų antspaudų ar DNR tuo, kad jo savybės laikui bėgant kinta sparčiausiai. Žmogaus balsas priklauso nuo jo savijautos, emocinės būsenos, taip pat jis gali būti specialiai keičiamas (siekiant pamėgdžioti kitą asmenį). Jo nepastovumas daro žmogaus balsą mažiau reikšminga ir patikima biometrine charakteristika, negu, pavyzdžiui, pirštų antspaudai, akies rainelė ar DNR. Kadangi kalbančiojo atpažinimas biometrijoje tampa labai svarbus, vienas iš pagrindinių tikslų, kuriant asmens atpažinimo pagal balsą sistemas, yra patikimų požymių, unikaliai atspindinčių asmenį, radimas, nes tai iki šiol dar nėra padaryta. Nors šiuo metu pasaulyje daugelyje biometrinių saugumo sistemų taikoma pirštų antspaudų technologija, dėl aukščiau paminėtų balso savybes ribojančių faktorių, tačiau tikimasi, kad balso technologijos ateityje paplis žymiai plačiau.

Toliau šiame skyriuje nagrinėsime kalbančiojo atpažinimo sistemas, jų tipus, klasifikaciją. Pateiksime ir aptarsime automatinės kalbančiojo atpažinimo sistemos struktūrą, jos elementus. Taip pat apžvelgsime kalbančiojo atpažinimo sistemų raidą bei atpažinimo problemas.

1.1. Kalbančiojo atpažinimo pagrindinės sąvokos

1.1.1. Kalbančiojo identifikavimas ir verifikavimas

Automatinis kalbančiojo atpažinimas gali būti padalintas į dvi pagrindines dalis: *kalbančiojo identifikavimą* ir *kalbančiojo verifikavimą* (Campbell 1997; Furui 1997).

Kalbančiojo verifikavimo užduotis, arba lyginimas „vieno su vienu“, tai yra atsakymas į klausimą, ar frazę pasakęs asmuo, kuris skelbiasi esąs, yra iš tikrųjų tas. Kalbančiojo verifikavimo metu nežinomo asmens balsas yra lyginamas su žinomo asmens, kuriuo pasiskelbė nežinomasis, balso etalonu. Šio lyginimo metu gautas panašumo laipsnis dar palyginamas su šiam žinomam asmeniui nustatytu slenksčiu. Jei šis panašumo laipsnis viršija žinomam asmeniui nustatytą slenkstį, tuomet priimamas sprendimas, kad tai yra tas asmuo, priešingu atveju nežinomas asmuo atmetamas kaip „svetimas“.

Kalbančiojo identifikavimas arba lyginimas „vieno su N “, dar savo ruožtu gali būti suskirstytas į *atviros* ir *uždaros aibės*. Vykdant *uždaros aibės*

kalbančiojo identifikavimą, nežinomas kalbėtojas yra lyginamas su visais N atpažinimo sistemoje užfiksuotais kalbėtojais. Rastas artimiausias kalbėtojas grąžinamas kaip atpažinimo rezultatas. Uždaros aibės kalbančiojo identifikavimo sistema daro priverstinį sprendimą, paprasčiausiai surasdama labiausiai atitinkantį kalbėtoją iš saugojamų bazėje, nesvarbu, koks prastas būtų tas atitikimas.

Jeigu yra galimybė, kad nei vienas iš registruotų bazėje kalbėtojų neatitinka nežinomo kalbėtojo, tai jau *atviros aibės* uždavinys. Atviros aibės kalbančiojo identifikavimo sistema turi turėti ir tam tikrą iš anksto nustatytą slenkstį, kad panašumo laipsnis tarp nežinomo kalbėtojo ir rasto artimiausio bazėje būtų didesnis už šį slenkstį. Taigi, kalbančiojo verifikavimas gali būti traktuojamas kaip atviros aibės kalbančiojo identifikavimas, kuomet sistemoje registruotas tik vienas kalbėtojas ($N = 1$) .

1.1.2. Nuo teksto priklausantis ir nepriklausantis kalbančiojo atpažinimas

Kalbančiojo atpažinimas savo ruožtu dar gali būti suskirstytas į priklausantį nuo pasakyto teksto ir nepriklausantį. Pirmuoju atveju į atpažinimo sistemą pateikiama frazė yra iš anksto žinoma. Antruoju atveju gali būti ištarta bet kokia frazė (fonetiniu turiniu panaši į pateiktas mokymo metu). Šiuo atveju atpažinimo sistema nustato bendrus asmens balso trakto požymius.

Nepriklausančiame nuo ištarto teksto kalbančiojo atpažinime paprastai naudojamos gerokai ilgesnės frazės nei priklausančiame, kadangi pirmuoju atveju, siekiant didesnio atpažinimo tikslumo, abi skirtingos frazės fonetiniu turiniu turi būti panašios, kad jas būtų galima efektyviai palyginti.

Bendru atveju priklausančios nuo ištarto teksto kalbančiojo atpažinimo sistemos veikia tiksliau nei nepriklausomos, kadangi atpažįstama frazė ir balsas. Kalbos atpažintuvas gali būti panaudotas frazės, pateiktos vartotojui, atpažinimui. Tai vadinama frazės verifikavimu ir gali būti efektyviai sujungta su asmens verifikavimo sistema (Li *et al.* 2000).

Priklausančiame nuo ištarto teksto kalbančiojo verifikavime gali būti naudojama ta pati raktinė frazė arba jos gali būti keičiamos kiekvienos sesijos metu. Antruoju atveju sistema atsitiktinai parenka raktinę frazę ir ją pateikia ištarti vartotojui. Raktinės frazės gali būti saugojamos kaip atskiri žodžiai ar sakiniai, kitas variantas – jos gali būti tiesiogiai suformuojamos iš atskirų žodžių ar kitų kalbos vienetų (garsų ar dvigarsių).

1.2. Kalbančiojo atpažinimo sistemų tipai

Kalbančiojo atpažinimą galima būtų suskirstyti tris pagrindinius tipus:

- Adityvinis kalbančiojo atpažinimas.
- Pusiau automatinės kalbančiojo atpažinimo sistemos.
- Automatinės kalbančiojo atpažinimo sistemos.

1.2.1. Adityvinis kalbančiojo atpažinimas

Adityvinį kalbančiojo atpažinimą mes atliekame kiekvieną dieną. Klausydami mes atpažįstame savo draugus, pažįstamus. Netgi išgirdę trumpą kalbos fragmentą, mes galime apibūdinti žmogaus lytį ar apytikslį amžių bei kitus kalbėtojo požymius. Adityvinis kalbančiojo atpažinimas plačiai taikomas atliekant fonoskopines teismo ekspertizes. Tačiau adityvinis kalbančiojo atpažinimas yra subjektyvus, neretai skirtingų klausytojų išvados gali skirtis. Be to, kuo ilgesnis laiko tarpas tarp dviejų išgirstų frazių, kurias reikia palyginti, tuo žmogaus galimybės atpažinti asmenį prastėja (Kerstholt *et al.* 2003). Dėl šių priežasčių šis atpažinimo būdas nėra laikomas labai patikimu atliekant teisminius tyrimus.

Siekiant palyginti žmogaus ir kompiuterio sugebėjimus atpažinti kalbantįjį, buvo atlikti specialūs tyrimai (Liu *et al.* 1997; Schmidt-Nielsen, Crystal 2000). Schmidt–Nielsen bei Crystal (Schmidt-Nielsen, Crystal 2000) atliko eksperimentus, t. y. apie 50 000 atpažinimo testų, kai 65 klausytojai sugrupuoti į grupes po 8 žmones. Šie rezultatai buvo palyginti su kompiuterių atpažinimo rezultatais. Buvo pastebėta, kad atskiri klausytojai pateikdavo labai skirtingus atpažinimo rezultatus, be to, jie nustatydavo labai skirtingus atpažinimo slenksčius. Kitais žodžiais tariant, balansas tarp klaidingo „savojo“ atmetimo ir klaidingo „svetimo“ priėmimo klaidų labai priklausė nuo klausytojo. Taip pat buvo pastebėta (Schmidt-Nielsen, Crystal 2000), kad esant įvairių foninių triukšmų, skirtingų įrašymo kanalų bei kitų triukšmų šaltinių iškraipytiems kalbos signalams, klausytojų atpažinimo rezultatai buvo geresni nei kompiuterių. Esant panašioms įrašymo sąlygoms bei neiškraipytiems kalbos signalams, gauti atpažinimo rezultatai buvo panašūs tiek klausytojų, tiek ir kompiuterių.

Tačiau šie tyrimai buvo atlikti naudojant NIST 1998 balsų bazę. Per tą laiką jau gerokai patobulėjo kompiuteriniai algoritmai ir šie rezultatai gali būti laikomi pasenusiais. Pastaraisiais metais sukurtos sistemos, naudojančios tiek žemo, tiek ir aukšto lygio kalbančiojo požymius (Reynolds *et al.* 2003; Campbell *et al.* 2003). Prie aukšto lygio požymių galima būtų priskirti prozodinę statistiką, idiolektą, tarties modeliavimą ir t. t. Derinant ir apjungiant skirtingus požymius, gaunami tikslesni atpažinimo rezultatai.

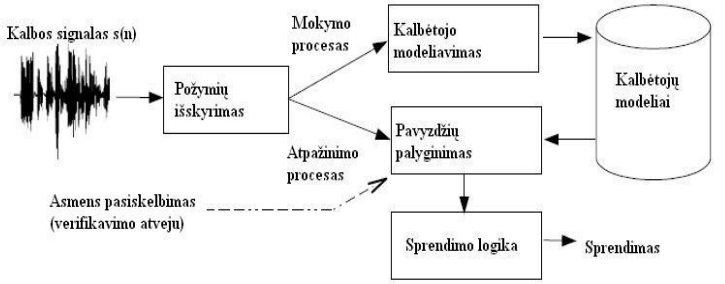
1.2.2. Pusiau automatinės kalbančiojo atpažinimo sistemos

Kai kuriais atvejais, pavyzdžiui, atliekant fonoskopines teismo ekspertizes, yra naudojamos pusiau automatinės kalbančiojo atpažinimo sistemos. Tai yra audityvinio atpažinimo ir automatinių atpažinimo sistemų hibridas. Dar Rose teigė (Rose 2002), kad atliekant teisminius tyrimus ir siekiant didžiausio tikslumo, reikia apjungti skirtingus metodus. Taip pat asmuo, atliekantis atpažinimą, turi turėti tam tikrus lingvistikos pagrindus. Reikia, kad du lyginami kalbos pavyzdžiai turėtų tuos pačius lingvistinius parametrus. Tuo tikslu reikia tiksliai išrinkti lyginamus kalbos signalų pavyzdžius. Šiuo atveju lyginamų kalbos vienetų išrinkimas arba segmentavimas atliekamas rankiniu būdu, tuo metu klausant tuos kalbos signalų fragmentus bei naudojant vizualią informaciją (signalogramas bei jų spektrogramas). Signalo fragmentų klausymas yra labai naudingas segmentacijai. Tačiau galutinė analizė turi būti atlikta skaičiuojant ir lyginant segmentuotų bei lyginamų kalbos signalų parametrus, naudojant statistinę analizę ar pan.

1.2.3. Automatinės kalbančiojo atpažinimo sistemos

Automatinė kalbančiojo atpažinimo sistema – tai aparatinė ar programinė įranga, kuriai pateikus asmens ištartos frazės signalą, gaunamas atsakymas, tas ar ne tas (verifikavimo atveju) arba kuris (identifikavimo atveju) asmuo ištarė šią frazę. 1.1 paveiksle parodyta apibendrinta automatinės kalbančiojo atpažinimo sistemos schema. Ši sistema vykdo du procesus: *mokymo* ir *atpažinimo*. Mokymo metu būna sukuriami kalbėtojų modeliai, kurie po to būna sistemoje saugojami. Atpažinimo proceso metu padavus į sistemą nežinomo asmens kalbos signalą, sistema padaro sprendimą apie jo tapatybę.

Tiek mokymo, tiek ir atpažinimo proceso metu pradžioje yra vykdomas *požymių išskyrimas*, kurio tikslas pateiktas ištartos frazės signalo reikšmes paversti tam tikrais požymiais arba požymių vektoriais, atspindinčiais identifikuojamo asmens individualybę. Nauji požymių vektoriai turi mažesnę dimensiją nei pradiniai (kalbos signalo reikšmės). Požymių vektorių išskyrimas taip pat reikalingas ir tolesniam skaičiavimų sumažinimui. Požymių vektoriai – yra stabilesnis, patikimesnis, atsparesnis bei kompaktiškesnis pirminio kalbos signalo aprašymas.



1.1 pav. Automatinės kalbančiojo atpažinimo sistemos struktūra

Fig. 1.1. Structure of an automatic speaker recognition system

Kalbos signalo požymiai, naudojami asmens atpažinimui, turėtų tenkinti tokius reikalavimus (Rose 2002):

1. Didelis skirtumas tarp skirtingų kalbėtojų.
2. Mažas skirtumas tam pačiam kalbėtojui.
3. Lengvai skaičiuojami.
4. Atsparūs mėgdžiojimui ir balso maskavimui.
5. Atsparūs triukšmams ir iškraipymams.
6. Maksimaliai nepriklausomi nuo kitų požymių (t. y. nekoreliuoti tarp savęs).

Būtų galima išskirti tokius požymių tipus, naudojamus kalbančiojo asmens atpažinime (Kinnunen 2005):

- spektriniai požymiai;
- dinaminiai požymiai;
- žadinimo šaltinio požymiai;
- suprasegmentiniai požymiai;
- aukšto lygio požymiai.

Spektriniai požymiai atvaizduoja mažos trukmės kalbos signalo spektrą, jie daugiau atspindi fizines balso trakto charakteristikas. Čia naudojami TPM parametrai, TPMK, MSKK, formantės, spektrinės poros ir t. t.

Dinaminiai požymiai parodo spektrinių ar kitų požymių kitimą laike. Čia dažniausiai naudojami delta-, delta-delta požymiai, moduliacijos dažniai, vektorių autoregresijos koeficientai.

Žadinimo šaltinio požymiai susiję su žadinimo signalu. Čia naudojamas žadinimo signalo pagrindinis dažnis – F0, balsaskylės impulso forma ir t. t.

Suprasegmentiniai požymiai apima ilgesnius laiko intervalus. Čia gali būti naudojami žadinimo signalo pagrindinio dažnio kontūrai, intensyvumo kontūrai, mikroprozodija.

Aukšto lygio požymiai remiasi simboliniu informacijos tipu, kaip pvz. charakteringas žodžių naudojimas. Čia gali būti naudojami tokie požymiai, kaip išskirtinis žodžių naudojimas, tartis.

Mokymo metu, naudojant požymių vektorius, yra sukuriamas kalbėtojo modelis. Šis modelis turi būti būdingas tam kalbėtojui ir geriausiai atitikti jo ištartos frazės požymių vektorius: ne vien tik tuos, kurie buvo pateikti mokymui, bet ir tuos, kurie bus pateikti atpažinimo procese.

Yra du būdai, kaip vertinti kalbėtojo požymių vektorių pasiskirstymą: parametrinis (stochastinis) ir neparametrinis (etaloninis) (Campbell 1997; Duda *et al.* 2000; Fukunaga 1990).

Atpažinimo metu taip pat skaičiuojami požymių vektoriai iš nežinomo asmens ištarto kalbos signalo. Toliau, požymių palyginimo metu, skaičiuojamas panašumo matas tarp nežinomo asmens požymių vektorių ir kiekvieno iš saugojamų kalbėtojų modelių (identifikavimo atveju) arba vieno konkretaus asmens modelio (verifikavimo atveju). Panašumo matas priklauso nuo kalbėtojų sukurtų modelių tipo.

Sprendimo priėmimo metu, sprendimo modulis pagal panašumo kriterijų padaro galutinį sprendimą apie kalbančiojo tapatybę (kai kada su tam tikru pasikliovimo intervalu (Gish, Schmidt 1994; Huggins, Grieco 2002)). Sprendimo tipas priklauso nuo iškeltos užduoties. Kalbančiojo *verifikavimo* atveju priimamas binarinis sprendimas, tas ar ne tas asmuo. Kalbančiojo *identifikavimo* atveju yra dvi galimybės: jei yra *uždaros aibės* identifikavimas, sprendimas yra asmuo, kurio modelis labiausiai atitinka nežinomo asmens požymių vektorius. *Atviros aibės* identifikavimo atveju dar yra galimas papildomas sistemos sprendimas, kad nei vienas iš registruotų kalbėtojų neatitinka nežinomo kalbėtojo.

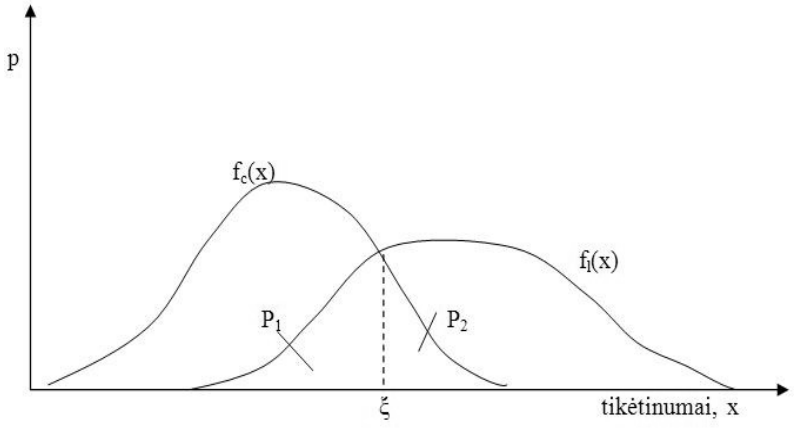
Praktikoje dažnai pasitaiko, kad sistemos rastas panašumo matas ne tiesiogiai lyginamas su slenksčiu, bet dar normalizuojamas (Furui 1997). Taip atsitinka, kadangi gali skirtis garso įrašymo sąlygos mokymo ir atpažinimo metu. Dėl to atsiranda papildomi iškraipymai ir gali būti taip, kad panašumo matai, gauti atpažinimo metu, gali labai skirtis nuo gautų mokymo metu. Tam tikslui yra pasiūlyta atitinkamų metodų (Bimbot *et al.* 2000).

1.3. Biometrinių sistemų darbingumo įvertinimas

1.3.1. Intraindividualūs ir interindividualūs pasiskirstymai

Mokymo metu, lyginant to paties asmens ištartas frazes su jo sukurtu modeliu, bus gauta aibė intraindividualių panašumo įverčių (tikėtinumų), kurių skaičius bus lygus lyginimui skirtų frazių skaičiui. Lyginant asmenų ištartas

frazes su kito asmens sukurtu modeliu, tam asmeniui gaunama aibė interindividualių panašumo įverčių (tikėtinumų), kurių skaičius taip pat bus lygus lyginimui skirtų frazių skaičiui. Interindividualių ir intraindividualių tikėtinumų pasiskirstymo grafikų pavyzdys parodytas 1.2 paveiksle (čia funkcija $f_c(x)$ vaizduoja interindividualių, o funkcija $f_l(x)$ vaizduoja intraindividualių tikėtinumų pasiskirstymą).



1.2 pav. Intraindividualių ir interindividualių panašumo įverčių pasiskirstymai

Fig. 1.2. Distributions of interindividual and intraindividual similarity scores

Bendruoju atveju intraindividualūs tikėtinumai yra didesni už interindividualius tikėtinumus. Idealiu atveju šios kreivės neturėtų persikirsti (mažiausias tikėtinumas tarp to paties asmens ištartų lyginamų frazių turėtų būti didesnis už didžiausią tikėtinumą tarp to asmens ir kito asmens ištartų frazių). Tuomet atpažįstant asmenį pagal balsą klaidų nebūtų. Tačiau realiai dažnai jos persikerta, dėl to atsiranda atpažinimo klaidos, kurios gali būti dviejų tipų: klaida, kad „savas“ asmuo bus priimtas už „svetimą“ ir klaida, kad „svetimas“ asmuo bus priimtas už „savą“. Pirmojo tipo klaidos tikimybę, kad „savas“ asmuo bus priimtas už „svetimą“, t. y. „savas“ asmuo bus atmestas, esant užduotam slenksčiui ξ, galima išreikšti:

$$P_1 = \int_{-\infty}^{\xi} f_l(x)dx, \qquad (1.1)$$

o antrojo tipo klaidos tikimybė, kad „svetimas“ asmuo bus priimtas už „savą“, t. y., kad „svetimas“ asmuo bus atpažintas, kaip „savas“, bus lygi:

$$P_2 = \int_{\xi}^{\infty} \mathrm{f}_c(x)dx\,. \tag{1.2}$$

Bendra klaidos tikimybė gali būti išreikšta pasinaudojus Bejeso formule:

$$P(\xi) = \alpha P_1(\xi) + (1-\alpha)P_2(\xi)\,, \tag{1.3}$$

čia α ir $(1–\alpha)$ atitinkamai „savo“ ir „svetimo“ balso pasirodymo tikimybės.

1.3.2. Biometrinių sistemų vertinimo charakteristikos

- Klaidingo priėmimo lygis KPL (angl. *FAR – false acceptance rate*) – tikimybė, kad sistema neteisingai paskelbia sėkmingą atitikimą tarp įėjimo pavyzdžio ir jį neatitinkančio pavyzdžio, saugojamo sistemos duomenų bazėje. Kitais žodžiais tariant, „svetimas“ asmuo priimamas kaip „savas“. Ši charakteristika nurodo neteisingų atitikimų tikimybę:

$$KPL = \frac{P}{R}\,, \tag{1.4}$$

čia P – klaidingų priėmimų (KP) skaičius, R – įėjimų skaičius.

- Klaidingo atmetimo lygis KAL (angl. *FRR – false reject rate*) – tikimybė, kad sistema neteisingai paskelbia neatitikimą tarp įėjimo pavyzdžio ir jį atitinkančio pavyzdžio, saugojamo sistemos duomenų bazėje. Kitais žodžiais tariant, „savas“ asmuo atmetamas kaip „svetimas“. Ši charakteristika nurodo atmestų teisingų atitikimų tikimybę:

$$KAL = \frac{Q}{W}\,, \tag{1.5}$$

čia Q – klaidingų atmetimų (KA) skaičius, W – įėjimų skaičius.

- Bendras klaidų lygis BKL (angl. *TER – total error rate*). Tai yra bendras KPL ir KAL klaidų lygis:

$$BKL = \frac{P+Q}{R+W}\,. \tag{1.6}$$

- Imtuvo darbinė charakteristika ROC (angl. – *receiver operating characteristic*). Tapatinimo algoritmas priima sprendimą naudodamas tam tikrus parametrus, pvz. slenkstį. Biometrinėse sistemose KPL ir KAL gali būti suderintos keičiant šį parametrą. ROC grafikai braižomi keliais būdais: vienas būdas atidedant KPL ir KAL reikšmes, kurios gaunamos keičiant atitinkamus parametrus, pvz. slenkstį. Kitas būdas, abscisių ašyje atidedant KPL reikšmes, ordinačių ašyje – teisingų atpažinimų tikimybę, procentais.

Bendras variantas yra DET (angl. – *detect error trade-off*) kreivė, kuri gaunama ant vienos ašies (ordinačių) atidėjus KAL, ant kitos (abscisių) – KPL (Martin *et al.* 1997).

- Lygių klaidų lygis LKL (angl. *EER – equal error rate*). Tai toks klaidos lygis, kuomet klaidingo „savojo“ atmetimo ir klaidingo „svetimo“ priėmimo klaidų tikimybės yra lygios, t. y. klaida, kuomet KPL=KAL. Nustatant LKL naudojamos ROC arba DET kreivės, kadangi jose aiškiai parodytas KPL bei KAL reikšmių kitimas kintant parametrams, pvz. slenksčiui. LKL parametras naudojamas tada, kai reikia greitai palyginti dvi sistemas. LKL reikšmę paprasčiausiai galima gauti iš ROC kreivės, imant tašką, kur KAL kreivė susikerta su KPL kreive. Kuo mažesnė sistemos LKL reikšmė, tuo tiksliau dirba sistema.
- Registravimo sutrikimų lygis FTE (angl. – *failure to enroll*). Tai yra procentai įejimo duomenų, kurie laikomi netinkamais ir į sistemą nepriimami. Taip atsitinka tada, kai iš sensoriaus gauti duomenys yra laikomi netinkamais ar per prastos kokybės.
- Išskyrimo sutrikimų lygis FTC (angl. – *failure to capture*). Kai įėjimo duomenys yra tinkami, tai parodo tikimybę, kad sistema nesugebės iš šių duomenų išskirti reikiamų biometrinių charakteristikų (požymių).
- Etalonų talpa: duomenų rinkinių, kurie gali būti įvesti į sistemą, maksimalus skaičius.

1.4. Kalbančiojo atpažinimo sistemų raida

Kalbos ir kalbančiojo atpažinimo sistemoms jau virš 50 metų. Per tą laiką buvo sukurta daug įvairių metodikų, kurios buvo nuolat tobulinamos, kol pasiekė šiuolaikinį lygį. Kuriant automatines kalbos/kalbančiojo atpažinimo sistemas buvo susiduriama su daugeliu problemų, tokių kaip atpažinimas esant nepalankioms sąlygoms, tinkamų požymių suradimas ir t. t. Nors šios problemos buvo sprendžiamos daugelį metų, kol kas nemaža dalis iš jų nėra iki galo išspręsta ir iki šių dienų.

Kalbančiojo atpažinimo sistemos atsirado dešimtmečiu vėliau nei kalbos atpažinimo sistemos, tačiau jos turi nemažai bendrų bruožų, jų raida tarpusavyje susijusi.

Toliau trumpai apžvelgsime kalbančiojo atpažinimo sistemų raidą, per pastaruosius 50 metų (Furui 2005).

1960–1970 metai.

Pirmosios pastangos sukurti automatinio kalbančiojo atpažinimo sistemas buvo 1960 metais. Pruzansky iš Bell Labs buvo vienas iš pirmųjų, pradėjęs

tyrimus naudojant filtrų rinkinius bei skaitmeninių spektrogramų lyginimą tam, kad būtų galima rasti panašumo matą (Pruzansky 1963). Doddingtonas iš Texas Instruments (TI) filtrų rinkinius pakeitė formančių analize (Doddington 1971). Viena iš didžiausių tuometinių problemų buvo didelis požymių skirtumas tam pačiam kalbėtojui. Šiuos klausimus sprendė ar nagrinėjo Endress (Endress 1971) ir Furui (Furui 1974).

Kuriant nepriklausomus nuo pasakyto teksto kalbančiojo atpažinimo metodus buvo naudojami įvairūs parametrai, nepriklausomi nuo fonetinio turinio. Šie parametrai buvo gaunami vidurkinant pakankamai ilgos trukmės frazės pradinius parametrus, arba gaunant statistinius bei prognozuojamus parametrus. Buvo naudojami tokie parametrai kaip suvidurkinta autokoreliacinė funkcija (Bricker *et al.* 1971), spektro bei pagrindinio tono histogramos (Beek *et al.* 1977), tiesinės prognozės modelio koeficientai (Sambur 1972) ar ilgos trukmės vidurkinis spektras (Furui *et al.* 1972).

Kadangi naudojant nepriklausomus nuo ištarto teksto kalbančiojo atpažinimo metodus nebuvo gaunami labai geri rezultatai, buvo kuriami ir priklausantys nuo ištarto teksto metodai (Atal 1974; Furui 1981; Rosenberg, Sambur 1975), kuriais buvo galima lyginti panašaus fonetinio turinio požymius. Naudojant šiuos metodus buvo gaunami žymiai geresni atpažinimo rezultatai.

TI sukūrė pirmąją pilnai automatizuotą kalbančiojo verifikavimo sistemą (Furui 2005). Spektro analizei buvo panaudoti skaitmeninių filtrų rinkiniai.

Bell Labs sukūrė eksperimentinę sistemą, skirtą dirbti su telefoninėmis linijomis. Furui pasiūlė naudoti kepstro koeficientus kartu su jų pirmais bei antrais polinominiais koeficientais (Furui 1981). Šie kalbos signalo kadro požymiai buvo skirti patikimumo didinimui, naudojant telefoninius kanalus. Kepstriniai požymiai vėliau tapo standartu ne tik asmens, bet ir kalbos atpažinime.

1980-ieji metai.

Kaip alternatyva pavyzdžių palyginimo metodams, naudojamiems priklausomame nuo ištarto teksto kalbančiojo atpažinime, buvo pradėti naudoti paslėptųjų Markovo modelių (PMM) metodai, tokie patys, kaip ir kalbos atpažinime. PMM architektūra pagrįstos kalbančiojo atpažinimo sistemos naudojo kalbančiųjų modelius, gautus iš sakinių, atskirų žodžių ar net fonemų. Dažniausiai buvo naudojamos keleto žodžių frazės, bei kiekvienam žodžiui buvo kuriami modeliai, kurie buvo kombinuojami sakinio lygyje, atsižvelgiant į sakinio lygio gramatiką (Naik *et al.* 1989).

Buvo kuriami ir nuo teksto nepriklausomi kalbančiojo atpažinimo metodai. Tam tikslui buvo kuriami neparametriniai bei parametriniai tikimybiniai modeliai. Kaip neparametrinis modelis buvo sukurtas vektorinio kvantavimo (VK) metodas (Soong *et al.* 1987), kuomet tam tikras trumpalaikių požymių vektorių skaičius buvo atvaizduojamas mažesniu požymių vektorių rinkiniu,

vadinamu kodine knyga. Kaip parametrinis modelis, buvo sukurtas PMM. Pritz (Pritz 1982) pasiūlė naudoti ergodinį PMM (kai leidžiami visi perėjimai tarp būsenų). Žodis buvo traktuojamas kaip perėjimų seka akustinių požymių erdvėje per 5 būsenų PMM. Tishby (Tishby 1991) toliau plėtojo šią idėją ir sudarė 8 būsenų ergodinį autoregresinį PMM, atvaizduojamą tolydinio tikimybinio tankio funkcijomis su nuo 2 iki 8 komponenčių mišiniuose, atitinkančiuose kiekvieną būseną. Rose (Rose, Reynolds 1990) pasiūlė naudoti vienos būsenos PMM, kuris dabar labai populiarus ir vadinamas Gauso mišinių modeliu (GMM).

1990 metais labiausiai buvo nagrinėjamos robastinio (patikimo) kalbančiojo atpažinimo problemos. Matsui (Matsui, Furui 1992) lygino vektorinio kvantavimo metodą su diskretinio/tolydinio tankio PMM metodais. Buvo nustatyta, kad tolydinio tankio ergodinis PMM patikimumo požiūriu gerokai pralenkia diskretinio tankio ergodinį PMM. Kai yra pakankamai mokymo duomenų, ergodinis tolydinio tankio PMM patikimumo požiūriu veikia panašiai kaip ir VK metodas. Jie nustatė kalbančiojo identifikavimo kokybės laipsnį, naudojant tolydinio tankio PMM kaip funkciją nuo būsenų bei mišinių skaičiaus. Buvo parodyta, kad tai nepriklauso nuo būsenų skaičiaus, bet labai priklauso nuo bendro mišinių skaičiaus. Tai reiškia, kad naudojant PMM nepriklausomame nuo teksto kalbančiojo atpažinime, perėjimai tarp skirtingų būsenų yra bereikšmiai, dėl to GMM metodas gali pasiekti tą patį tikslumą kaip ir daugelio būsenų ergodinis PMM.

Matsui (Matsui, Furui 1993) pasiūlė pasufleruoto teksto (angl. – *text-prompted*) kalbančiojo atpažinimo metodą. Šiame metode kiekvienu sistemos panaudojimo atveju raktiniai sakiniai yra visiškai keičiami. Sistema nustato, ar registruotas kalbėtojas ištarė pasufleruotą tekstą. Kadangi žodynas yra neribotas, galimi apsimetėliai negali žinoti teksto, kuris jiems bus pasufleruotas ištarimui. Šis metodas ne tik tiksliai atpažįsta kalbantįjį, bet ir atmeta frazes, kurios skiriasi nuo pasufleruotų, nors gali būti ištartos registruoto kalbėtojo.

Taip pat buvo sprendžiamos ir tokios problemos, kaip intraindividualių (t. y. tarp to paties asmens) kitimų normalizavimas. Kitais žodžiais tariant, atsiranda tam tikras atstumas tarp dviejų to paties asmens ištartų frazių. Šios problemos atsiranda dėl įvairių priežasčių: skirtingos įrašymo sąlygos, perdavimo kanalo skirtumai, triukšmų lygis ir pan. Be to kalbėtojai negali du kartus visiškai vienodai ištarti to paties teksto. Buvo pasiūlyti tikėtinumo santykio bei aposteriorinėmis tikimybėmis pagrįsti metodai (Higgins *et al.* 1991; Matsui, Furui 1994).

2000 metais taipogi buvo sprendžiamos įverčio normalizavimo problemos. Be to, šiuo metu buvo pasiūlyti nauji aukšto lygio požymiai, naudojami kalbančiojo verifikavime: žodžio idiolektas, tartis, kalbos garsų naudojimas, prozodija ir t. t. (Campbell *et al.* 2007).

1.5. Kalbančiojo atpažinimo problemos

Asmens atpažinimas pagal balsą nėra labai patikima biometrijos rūšis, ji tikslumu ir patikimumu nusileidžia kitoms biometrijos rūšims, tokioms kaip asmens atpažinimas pagal jo pirštų antspaudus, akies rainelę ir t. t. Dalis šių problemų kyla dėl žmogaus anatominių bei fiziologinių savybių, kita dalis dėl techninės įrangos netobulumo. Toliau trumpai aptarsime dalį problemų.

Kaip žinoma, žmogaus balsas nėra visą laiką pastovus ir yra įtakojamas daugelio veiksnių. Yra žinoma, kad tam tikri psichologiniai veiksniai (depresija, susinervinimas, stresas, baimė ir t. t.), taip pat įvairūs sveikatos sutrikimai (ligos, galvos skausmas ir t. t.) veikia žmogaus balsą. Įvairių stimuliatorių vartojimas (narkotikai, alkoholis, rūkymas ir t. t.) taipogi daro tam didelę įtaką. Yra žinoma, kad net tą pačią dieną su ta pačia įranga kelis kartus įrašius vieno asmens balsą ir jį palyginus**,** jis visada kažkiek skirsis, o kartais tarp jų gali būti ir nemažas skirtumas. Taip pat balso pokyčiams įtakos turi ir tokie fiziologiniai veiksniai, kaip amžiaus kitimas, antsvorio atsiradimas ar išnykimas ir t. t. Be to vienus kalbėtojus gali būti sunkiau modeliuoti nei kitus (Doddington *et al.* 1998).

Dalis problemų gali būti susiję ir su *mėgdžiojimu* ar *balso maskavimu*. Mėgdžiojimas – tai yra, kai vienas asmuo bando keisti savo balsą taip, kad jo balsas būtų panašus į kito asmens. Balso maskavimas – tai balso keitimas siekiant, kad jis taptų nepanašus į savojo. Balso maskavimas dažniausiai sutinkamas kriminalistikoje. Pavyzdžiui, skambindamas nusikaltėlis gali specialiai keisti savo balsą, kad jo neatpažintų, pavyzdžiui, kalbėti užspaudęs nosį, įsikandęs kokį nors daiktą, ar kalbėti šnabždėdamas ir t. t. Dalis tyrimų buvo atlikta šioje srityje ir buvo ieškoma požymių, atspariausių balso maskavimui (Majewski, Mazur-Majewska 1999), tačiau ši problema nėra iki galo išspręsta.

Mėgdžiojimas ir balso maskavimas labai įtakoja atpažinimo tikslumą ir gali ženkliai pabloginti atpažinimo rezultatus. Tiek mėgdžiojant, tiek maskuojant balsą yra keičiami kai kurie akustiniai kalbėtojo parametrai ir jie nesutampa su natūraliai kalbančio kalbėtojo akustiniais parametrais.

Atpažinimo tikslumą gali įtakoti ir techniniai klaidų šaltiniai. Galima būtų išskirti du tokius klaidų šaltinius: *akustinė aplinka* ir *perdavimo kelias*. Dėl akustinės aplinkos (pridedančios adityvinius triukšmus) – tai dėl foninio triukšmo, aplinkos akustinių savybių, aido ir pan. atsirandančios klaidos. Įrašant balsą per mikrofoną prie signalo pridedamas aplinkos foninis triukšmas, pvz. automobilių garsas, šalia esančios įrangos (televizoriaus, radijo, šaldytuvo) skleidžiamas triukšmas, kitų žmonių kalba ir t. t. Reverbacija prideda pavėlintą originalaus signalo kopiją prie įrašomojo (Huang *et al.* 2001).

Kita dalis iškraipymų atsiranda dėl signalo perdavimo kelio (pridedančio multiplikatyvinius triukšmus). Tai yra mikrofono iškraipymai, įrašymo įrangos trukdžiai, ribota pralaidumo juosta, analoginio – skaitmeninio keitiklio (ASK) kvantavimo triukšmai, kalbos kodavimo iškraipymai.

Labai didelę įtaką garso įrašo kokybei turi mikrofono charakteristikos. Prastos kokybės mikrofonai labai keičia signalo spektrą, įvedami netiesinius kalbos signalo spektro iškraipymus. Tyrimais buvo parodyta, kad dėl prastos kokybės mikrofonų kalbos signalo spektre atsiranda papildomos netikros formantės, praplatėja formančių dažnių juostos, pats spektras ištiesinamas (Quatieri *et al.* 2000).

Įrašymo aparatūra gali būti paveikta kitų trikdžių, pvz. šalia esančių mobiliųjų telefonų spinduliuojamų radijo bangų, kurios dėl indukcijos sukelia įrašymo aparatūroje papildomus trikdžius. Taip pat signalas iškraipomas analoginio – skaitmeninio keitiklio (ASK), verčiant jį iš analoginės į skaitmeninę formą. Jei kalbos signalas perduodamas telefoniniu kanalu, jis papildomai suspaudžiamas kompresijos metodais be tikslaus atstatymo, dėl to atsiranda papildomi triukšmai ir iškraipymai. Kalbos kodavimas gali labai pabloginti atpažinimo sistemos veikimo rezultatus (Phythian *et al.* 1997).

Viena iš svarbiausių problemų asmens atpažinime pagal balsą yra *neatitinkančios sąlygos* (Li *et al.* 2001; Mammone *et al.* 1996; Reynolds 2002). Kitais žodžiais tariant, skiriasi garso įrašymo aplinkybės sistemos mokymo ir atpažinimo metu. Nepaisant intraindividualaus kitimo, atsiranda dar ir techninis neatitikimas, pavyzdžiui, aplinkos akustinis neatitikimas, foninio triukšmo (tipo ir kiekio) neatitikimas, mikrofono, bei įrašymo kokybės neatitikimas ir t. t. Nesudėtinga įsivaizduoti situaciją, kai, pavyzdžiui, sistemos mokymas atliekamas tylioje aplinkoje, panaudojant geros kokybės mikrofoną, tuo tarpu atpažinimas atliekamas triukšmingoje gatvėje, dar panaudojus prastesnės kokybės mikrofoną. Suprantama, kad tokiu atveju atpažinimo rezultatai bus labai prasti.

Kita didžiulė problema kalbančiojo atpažinimo sistemose, kad visi spektriniai kalbos signalo požymiai, labiau priklauso nuo kalbos nei nuo kalbančiojo. Iki šiol nėra rasta tokių kalbos signalo požymių, kurie vienareikšmiškai nusakytų asmenį ir nepriklausytų nuo kalbos. Kokius spektrinius požymius nepaimtumėme, jei, pavyzdžiui, lyginsime vieno asmens ištartą fonemą A su to paties asmens ištarta fonema O, tai atstumas tarp jų bus visada didesnis nei tarp to asmens fonemos A ir kito asmens fonemos A. Todėl kalbančiojo atpažinimo sistemose lyginant du asmenis, reikia tarpusavyje lyginti tuos pačius jų ištartus garsus.

1.6. Pirmojo skyriaus apibendrinimas

- Kalbančiojo atpažinimas pagal balsą yra nuolat besivystanti biometrijos rūšis, kuriai pasaulyje kiekvienais metais skiriama vis daugiau lėšų, kuriami nauji algoritmai ir tikimasi, kad ji po kiek laiko paplis žymiai plačiau, panašiai kaip šiuo metu asmens atpažinimas pagal jo piršų antspaudus.

- Ši biometrijos rūšis nereikalauja jokios brangios specialios įrangos ar infrastruktūros, reikalingas tik mikrofonas ir programinė įranga.

- Naudojant priklausomą nuo pasakyto teksto kalbančiojo atpažinimą galima pasiekti didesnį atpažinimo tikslumą nei nepriklausomą, tačiau ne visur jis gali būti pritaikytas.

- Lietuvoje šioje srityje darbų padaryta labai mažai. Šiuo metu Lietuvoje daugiau dirbama kalbos atpažinimo srityje.

- Iki šiol nėra rasta kalbos signalo požymių, vienareikšmiškai atspindinčių asmens tapatybę. Dėl tos priežasties šiuo metu nėra sukurta patikimo, nepriklausančio nuo pasakyto teksto, asmens atpažinimo metodo.

2

Kalbančiojo atpažinimo sistemų analizė

Šiame skyriuje smulkiau panagrinėsime kalbančiojo atpažinimo sistemas. Šis skyrius padalintas į šešias dalis: pirmojoje dalyje aptarsime kalbos signalų generavimą, modeliavimą, jų savybes. Antrojoje dalyje paanalizuosime pirminio kalbos signalų apdorojimo metodus, taikomus beveik visose kalbos ir asmens atpažinimo sistemose. Trečiojoje dalyje panagrinėsime dažniausiai naudojamas požymių sistemas, ketvirtojoje dalyje aptarsime kalbos signalų atskyrimo nuo triukšmo (segmentavimo) metodus, penkta dalis skirta kalbos signalų palyginimo metodų, naudojamų tiek priklausančiame tiek nepriklausančiame nuo ištarto teksto kalbančiojo atpažinime, analizei. Šeštojoje dalyje trumpai bus apžvelgtos kai kurios šiuolaikinės atpažinimo sistemos.

2.1. Kalbos signalų generavimas ir modeliavimas

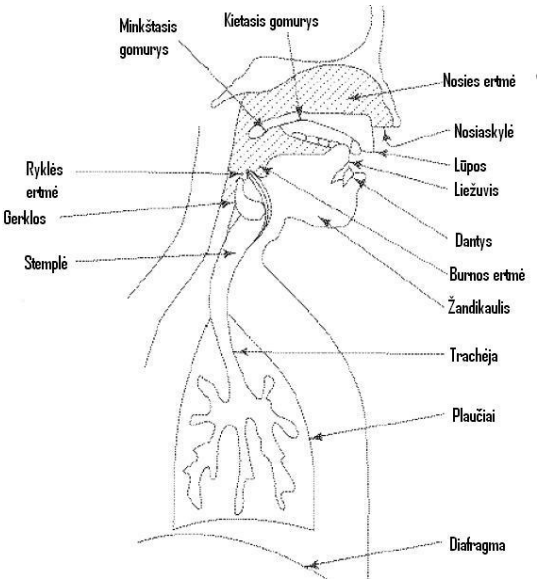
Kalba, be abejo, yra svarbiausia žmonių tarpusavio bendravimo priemonė. Kalbos signalas turi savyje informaciją apie pranešimo turinį, kalbantįjį asmenį, jo nuotaiką, emocinę būseną, lytį ir t. t. Nepriklausomame nuo ištarto teksto kalbančiojo atpažinime mus domina tik tos kalbos signalų savybės, kurios nusako

asmenį, visa kita informacija yra perteklinė. Priklausančiame nuo teksto kalbančiojo atpažinime taip pat svarbi yra informacija ir apie pranešimo turinį.

Kalbos generavimo principai yra gana neblogai ištirti ir pagal juos sukurti veikiantys modeliai, kurie plačiai taikomi praktikoje, pvz., mobiliojo ryšio aparate yra realizuotas veikiantis kalbos generavimo modelis.

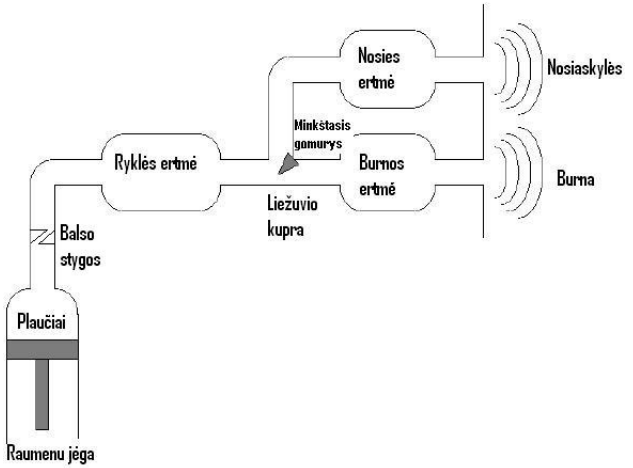
2.1.1. Kalbos signalų generavimas

Kalbos signalas yra akustinis oro molekulių spaudimas, susidarantis dėl žmogaus anatominių struktūrų judėjimo (Huang *et al.* 2001). Svarbiausi žmogaus kalbos signalo generavimo komponentai yra plaučiai, trachėja, gerklos su jų svarbiausia dalimi – balso stygomis, nosies ertmė, minkštasis bei kietasis gomurys, liežuvis, dantys bei lūpos (2.1 paveikslas). Visi šie komponentai, dar vadinami *artikuliatoriais*, juda keisdami savo padėtį, taip sukurdami skirtingus garsus. Kalbos generavimo procedūrą galima įsivaizduoti kaip akustinio filtravimo procesą, veikiantį iš plaučių sklindantį oro srautą. Šį akustinį filtrą sudaro trys pagrindinės ertmės: ryklės, nosies bei burnos (Deller *et al.* 2000). Šios ertmės bei artikuliatoriai sudaro taip vadinamą balso traktą. Šio trakto supaprastintas akustinis modelis pavaizduotas 2.2 paveiksle.



2.1 pav. Žmogaus balso trakto pjūvis

Fig. 2.1. Humans vocal tract



2.2 pav. Balso trakto supaprastintas akustinis modelis

Fig. 2.2. Simplified acoustic model of the vocal tract

Kalbos generavimo procesą galima būtų suskaidyti į tris etapus: šaltinio generavimas, balso trakto artikuliacija bei garso sklidimas nuo lūpų ir nosiaskylių (Furui 2001).

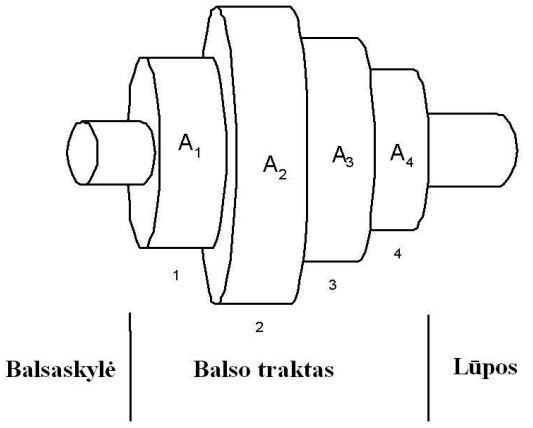
Šaltinio generuojamas signalas būna dviejų tipų: kvaziperiodinis bei aperiodinis. Vokalizuoti garsai (pvz. balsiai A, E ir t. t.) gaunami sklindant oro srautui iš plaučių ir virpant balso stygoms. Balso stygų virpėjimas yra kvaziperiodinis, o jų virpėjimo periodas vadinamas *pagrindiniu tonu*. Nevokalizuoti garsai (pvz. priebalsiai S, Š) gaunami, kai iš šaltinio sklinda turbulentinis, aperiodinis signalas (triukšmas) (Campbell 1997; Deller *et al.* 2000).

2.1.2. Balso trakto modeliavimas

Tam, kad sukurti automatinę kalbančiojo atpažinimo sistemą, visų pirma reikia sukurti efektyvų žmogaus kalbos signalo generavimo modelį bei vertinti šio modelio parametrus. Turint tokį modelį, mes galime nustatyti tam tikras kalbos signalo savybes (požymius), atitinkančias šį modelį. Tuomet pagal šias savybes būtų įmanoma nustatyti, ar du kalbos signalai atitinka tą patį modelį, kitaip tariant, ar priklauso tam pačiam asmeniui.

Paprastai modeliavimo procesas susideda iš dviejų etapų: žadinimo signalo modeliavimo ir balso trakto modeliavimo (Deller *et al.* 2000). Yra naudojama prielaida, kad žadinimo šaltinis ir balso traktas yra nepriklausomi (Campbell 1997; Deller *et al.* 2000). Vienas iš naudojamų modelių yra nuosekliai sujungtų vamzdžių modelis, kuris pagrįstas prielaida, kad kalbos signalas generuojamas

besikeičiant balso trakto formai. Kadangi formaliai aprašyti balso trakto kitimą laike yra labai sudėtinga, realiai šis modelis yra supaprastinamas ir vaizduojamas, kaip keletas nuosekliai sujungtų akustinių vamzdžių, su laike besikeičiančiais jų skerspjūvio plotais (Deller *et al.* 2000). Šių akustinių vamzdžių skerspjūvio plotas A_k ir ilgis L_k (2.3 paveikslas). Paprastumo dėlei priimama, kad šių vamzdžių ilgis yra vienodas. Naudojant daug labai trumpų vamzdžių, galima pagerinti šio modelio charakteristikas iki beveik tolygiai besikeičiančio skerspjūvio ploto, bet tuomet tampa sudėtingesnis pats modelis.

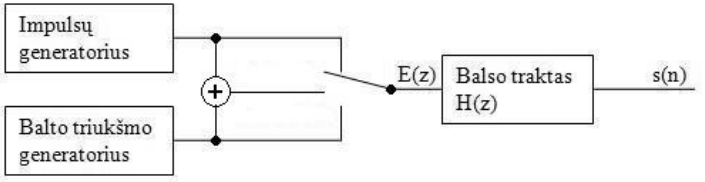


2.3 pav. Žmogaus balso trakto modelis

Fig. 2.3. Model of human‘s vocal tract

Yra naudojamas bendresnis diskretinio laiko balso trakto modelis, dar vadinamas *šaltinio – filtro* modeliu (Deller *et al.* 2000). Šiame modelyje šaltinis yra periodinė impulsų seka arba baltas triukšmas, arba jų kombinacija.

Šaltinio – filtro modelis pagrįstas ta prielaida, kad visus žmogaus ištartus garsus galima suskirstyti į tris kategorijas: vokalizuotus, nevokalizuotus bei jų kombinacija. Vokalizuoti garsai (pvz. balsiai A, E, I, O ir t. t.) gaunami, kai periodinis žadinimo signalas filtruojamas balso trakto filtru. Nevokalizuoti garsai (pvz. duslieji priebalsiai S, Š) gaunami, kai baltas triukšmas filtruojamas balso trakto filtru. 2.4 paveiksle pateiktoje schemoje $E(z)$ tai žadinimo signalo funkcija, $H(z)$ – perdavimo (balso trakto) funkcija, $s(n)$ – sugeneruotas kalbos signalas (Deller *et al.* 2000). Taigi, į balso traktą mes galime žiūrėti kaip į skaitmeninį filtrą, kuris filtruoja įėjimo (žadinimo) signalą. Pagal skaitmeninių filtrų teoriją, turėdami filtro išėjimą, mes galime apskaičiuoti skaitmeninio filtro parametrus.



2.4 pav. Šaltinio – filtro modelis

Fig. 2.4. Source – filter model

2.2. Pirminis kalbos signalų apdorojimas

Šioje dalyje bus aptarti kalbos signalų apdorojimo metodai. Pirminį kalbos signalų apdorojimą sudaro trys etapai: *pradinė filtracija*, *signalo dalijimas į kadrus* ir *lango funkcijos* taikymas. Šie trys etapai naudojami beveik visose kalbos ir kalbančiojo atpažinimo bei kitose sistemose.

2.2.1. Pradinė filtracija

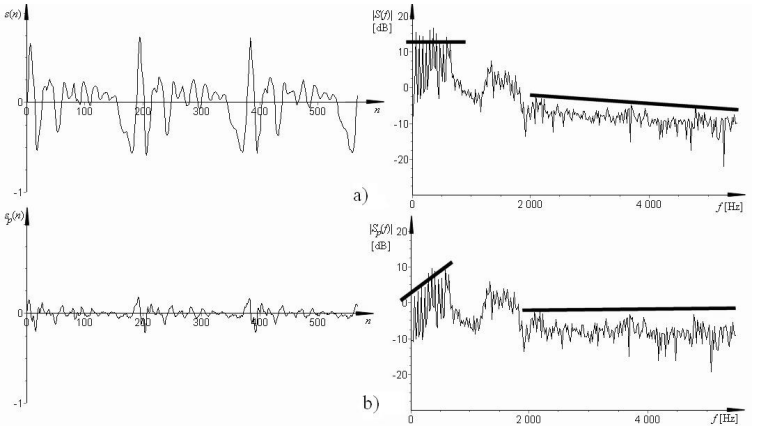
Kalbos signalo spektras daugiausia išsidėstęs žemų dažnių srityje, aukštesniuose dažniuose jo intensyvumas mažesnis ir krenta maždaug 6 dB į oktavą (t. y. du kartus padidėjus dažniui, spektro amplitudė sumažėja maždaug 6 dB). Dėl šios priežasties, pvz. aukštesnės formantės, yra žymiai mažiau išreikštos nei žemesnės, nors jos taip pat turi savyje svarbią informaciją apie kalbą bei asmenį. *Pradinės filtracijos* (Rodman 1999) tikslas – iškelti aukštesnių dažnių spektro komponentes tam, kad būtų galima padidinti jų įtaką bei pagerinti naudojamų požymių kokybę. Tokiu būdu yra nuslopinamos žemesnių dažnių spektro komponentės ir taip spektras „išlyginamas“.

Laiko srityje pradinė filtracija atliekama panaudojus žemos eilės skaitmeninį RIR filtrą. Dažniausiai naudojamas pirmos eilės RIR filtras, apibrėžiamas kaip:

$$\widetilde{s}(n) = s(n) - \alpha s(n-1)\,, \tag{2.1}$$

čia $\widetilde{s}(n)$ yra filtruotas signalas, $s(n)$ yra pradinės diskretizuoto šnekos signalo reikšmės, α – koeficientas, apsprendžiantis šnekos signalo spektro išlyginimo laipsnį, jis yra parenkamas intervale $0{,}9 \le \alpha \le 1{,}0$. Šio filtro sistemos funkcija:

$$H(z) = 1 - \alpha z^{-1}\,. \tag{2.2}$$



2.5 pav. Pradinės filtracijos įtaka: a) originalus signalas ir jo spektras; b) tas pats signalas ir jo spektras po pradinės filtracijos

Fig. 2.5. Influence of pre-emphasis: a) original signal and spectrum; b) the same signal and spectrum after pre-emphasis

Šio filtro DACh yra tiesė, kurios pasvirimo kampas priklauso nuo α, taigi gauname, kad signalo filtracija laiko srityje atitinka jo spektro dauginimą iš pasvirusios tiesės dažnių srityje (2.5 paveikslas).

Dėl pradinės filtracijos pagerėja pvz. TPM charakteristikos.

2.2.2. Signalų dalijimas į kadrus

Dalijimas į kadrus yra labai svarbus etapas kuriant kalbos ir kalbančiojo atpažinimo sistemas. Spektriniai kalbos signalų požymiai yra išskiriami iš trumpų kalbos signalo intervalų – kadrų, kurių trukmė apie 20–25 ms. Taip elgiamasi todėl, nes daroma prielaida, kad per trumpą laiko intervalą žmogaus balso trakto parametrai nespėja pasikeisti (Hui-Ling 2002), t. y. trumpame laiko intervale žmogaus balso traktą galima aprašyti pastoviais parametrais. Dažniausiai šie kadrai persikloja, t. y. sekantis kadras prasideda nuo prieš tai buvusio kadro tam tikros dalies, tuomet tarp kadrų atsiranda tam tikra koreliacija. Atskiras kadras gali būti išreikštas:

$$s(j,n) = s(j \cdot (N - O) + n), \tag{2.3}$$

čia $s(n)$ – originalus signalas, $s(j,n)$ – j-tasis kadras, N – kadro ilgis atskaitomis, O – gretimų kadrų persiklojimo ilgis atskaitomis.

2.2.3. Lango funkcijos taikymas

Prieš tolimesnį apdorojimą, kalbos signalo kadrai yra dauginami iš tam tikros lango funkcijos $w(n)$.

Signalą po lango funkcijos galime išreikšti:

$$s_w(n) = s(n) \cdot w(n). \tag{2.4}$$

Tiesiog paėmus signalo kadrą tai yra tolygu jį padauginti iš stačiakampio lango funkcijos:

$$w(n) = \begin{cases} 1, \text{ kai } 1 \le n \le N; \\ 0, \text{ kitur.} \end{cases} \tag{2.5}$$

Kadangi stačiakampį frontą suformuoti reikia begalinio spektro harmonikų skaičiaus, todėl stačiakampis langas turi labai prastas spektrines charakteristikas, t. y. toks langas labai iškraipo kalbos signalo spektrą. Dėl to parenkamos langų funkcijos, kurios ties lango pradžia ir pabaiga artėja prie nulio.

Naudojami įvairūs langai: Hemingo, Haningo, Gauso, Barleto (Bartlett) ir t. t.

Hemingo lango funkcija:

$$w(n) = \begin{cases} 0{,}54 - 0{,}46 \cdot \cos\left(2\pi \dfrac{n}{N}\right), 1 \le n \le N; \\ 0, \qquad\qquad\qquad \text{kitur.} \end{cases} \tag{2.6}$$

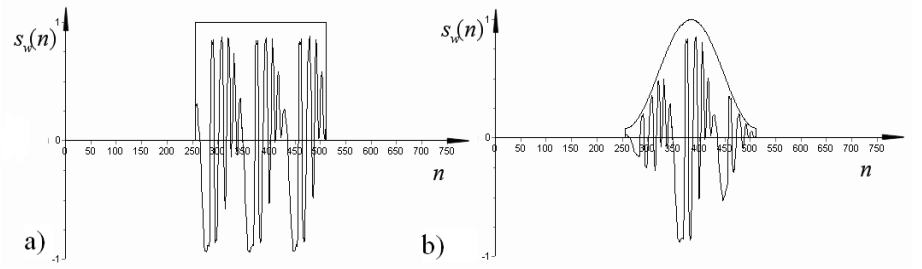
Haningo (Hanning) lango funkcija:

$$w(n) = \begin{cases} 0{,}5 - 0{,}5 \cdot \cos\left(2\pi \dfrac{n}{N}\right), 1 \le n \le N; \\ 0, \qquad\qquad\qquad \text{kitur.} \end{cases} \tag{2.7}$$

Gauso lango funkcija:

$$w(n) = \begin{cases} e^{-\frac{1}{2}\left(\alpha \frac{n}{N/2}\right)^2}, 1 \le n \le N; \\ 0, \qquad\qquad \text{kitur.} \end{cases} \tag{2.8}$$

Kalbos signalo kadro dauginimas iš stačiakampio bei Hemingo lango pavaizduotas 2.6 paveiksle.



2.6 pav. Signalo kadras, padaugintas iš: a) stačiakampio lango funkcijos; b) Hemingo lango funkcijos

Fig 2.6. Signal frame multiplied by: a) rectangular window; b) Hamming window

2.3. Kalbos signalų požymiai

Požymių išskyrimas yra vienas iš svarbiausių momentų kalbančiojo atpažinimo procese. Nuo jo labai priklauso kalbančiojo atpažinimo tikslumas. Geras požymių parinkimas ženkliai įtakoja atpažinimo rezultatus.

2.3.1. Energija

Signalo energija dažniausiai naudojama kalbos signalų segmentavimui, kai reikia atskirti kalbos signalą nuo triukšmo, kadangi kalbos signalo energija yra didesnė už triukšmo energiją. Fiksuoto ilgio diskretinio laiko signalo energija gali būti išreikšta:

$$E = \sum_{n=1}^{N_{\text{sum}}} s^2(n), \tag{2.9}$$

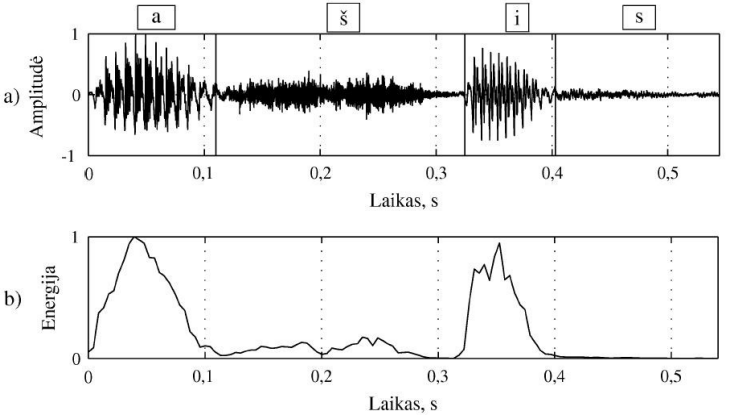
čia N_{sum} – signalo ilgis atskaitomis.

Diskretinio laiko signalui, padalintam į kadrus, *j*-tojo kadro trumpalaikė energija apskaičiuojama pagal formulę:

$$E(j) = \sum_{n=1}^{N} s^2(j,n), \tag{2.10}$$

čia N – signalo kadro ilgis atskaitomis.

Normuotos žodžio „ašis“ signalo laiko ir energijos diagramos pavaizduotos 2.7 paveiksle.



2.7 pav. Normuotos kalbos signalo diagramos (Tamulevičius 2008): a) laiko; b) energijos

Fig. 2.7. Diagrams of the speech signal (Tamulevičius 2008): a) time; b) energy

2.3.2. Nulių kirtimų dažnis

Nulių kirtimų dažnis (angl. – *zero crossing rate*) parodo, kiek kartų signalas kerta laiko ašį per sekundę. Taip pat šis požymis padeda atskirti įvairius balsius nuo triukšmo ar priebalsių ir yra dažnai naudojamas segmentavimui. Triukšmo nulių kirtimų dažnis yra daug didesnis nei, pavyzdžiui, balsių. Tačiau kai kurių priebalsių nulių kirtimų dažnis yra taip pat labai aukštas, panašus į triukšmo, dėl to juos sunkiau atskirti nuo triukšmo. Nulių kirtimų dažnį galima išreikšti:

$$Z(j) = \frac{1}{2}\sum_{n=1}^{N-1}\left|\text{sign}(s(j,n)) - \text{sign}(s(j,n+1))\right|, \qquad (2.11)$$

čia $N-$ j-tojo kadro ilgis atskaitomis, $j \in \{1, 2, ..., K\}$, $K-$ kadrų skaičius.

Funkcija $\text{sign}(s(n))$ gali būti išreikšta:

$$\text{sign}(s(n)) = \begin{cases} +1, \text{kai } s(n) > 0 \text{ arba } s(n) = 0 \text{ ir } s(n-1) > 0; \\ -1, \text{kai } s(n) < 0 \text{ arba } s(n) = 0 \text{ ir } s(n-1) < 0. \end{cases} \qquad (2.12)$$

2.3.3. Autokoreliacija

Koreliacija yra operacija, dažnai naudojama signalų apdorojime. Ji dažnai naudojama siekiant išskirti naudingą signalą iš triukšmo tam, kad būtų galima išskirti periodines signalų komponentes, sistemų identifikavimui ir t. t.

Jei turime dvi baigtinės energijos sekas, jų tarpusavio koreliaciją galime išreikšti:

$$r_{xy}(k) = \sum_{n=-\infty}^{\infty} x(n)y(n-k), k = 0,\pm 1,...,\pm\infty. \quad (2.13)$$

Tuo atveju, kai $y(n) = x(n)$, turime sekos $x(n)$ autokoreliaciją (Rodman 1999):

$$r_{xx}(k) = \sum_{n=-\infty}^{\infty} x(n)x(n-k) = \sum_{n=-\infty}^{\infty} x(n+k)x(n), k = 0,\pm 1,...,\pm\infty. \quad (2.14)$$

Jei $x(n)$ priežastinė N ilgio seka, tai:

$$r_{xx}(k) = \sum_{n=0}^{N-|k|-1} x(n)x(n-k) \ . \quad (2.15)$$

Vėlinimo parametras k kinta nuo 1 iki N–1 ir dažnai vadinamas autokoreliacinės funkcijos eile. k-tos eilės autokoreliacija parodo pradinio signalo ir perstumto per k vienetų, panašumą. Kalbos signalų apdorojime naudojami pirmieji 12÷32 autokoreliacinės funkcijos koeficientai. Paprastai autokoreliacinės funkcijos eilė apytiksliai parenkama $k=F_d+4$, kur F_d yra diskretizacijos dažnis, išreikštas kHz.

2.3.4. Tiesinės prognozės modelis

Sprendžiant kai kuriuos uždavinius tenka prognozuoti sistemos elgesį būsimais laiko momentais, kai turime sistemos išėjimus iki laiko momento n–1, tenka prognozuoti, koks bus sistemos išėjimas laiko momentu n. Tiesinės prognozės modelis (angl. – *linear predictive coding*) (Rodman 1999; Markel, Gray 1976) labai plačiai naudojamas sprendžiant kalbos bei kalbančiojo atpažinimo uždavinius. Šio modelio pagrindas yra kalbos signalo generavimo modelis „šaltinis – filtras“ (2.4 paveikslas). Sekanti kalbos signalo reikšmė gali būti apytiksliai apskaičiuota turint tam tikrą prieš tai buvusių signalo reikšmių skaičių p (autoregresinis modelis):

$$s[n] = \sum_{i=1}^{p} a_i s[n-i] + Gu[n], \quad (2.16)$$

čia a_i, i=1, 2,…, p – tiesinės prognozės modelio koeficientai, p – tiesinės prognozės modelio eilė, G – stiprinimo koeficientas, $u[n]$ – paklaidos signalas.

Taigi, prognozuotas signalas gali būti išreikštas:

$$\hat{s}[n] = \sum_{i=1}^{p} a_i s[n-i]. \tag{2.17}$$

Su prognoze susijusi klaida vadinama *prognozės klaida* arba dar dažnai vadinama *žadinimo signalu*:

$$e(n) = s[n] - \hat{s}[n] = s[n] - \sum_{i=1}^{p} a_i s[n-i]. \tag{2.18}$$

Tiesinės prognozės modelio koeficientai gali būti rasti sprendžiant matricinę lygtį:

$$\begin{pmatrix} r(0) & r(1) & \dots & r(p-1) \\ r(1) & r(0) & \dots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \dots & r(0) \end{pmatrix} \cdot \begin{pmatrix} a(1) \\ a(2) \\ \vdots \\ a(p) \end{pmatrix} = \begin{pmatrix} -r(1) \\ -r(2) \\ \vdots \\ -r(p) \end{pmatrix}, \tag{2.19}$$

čia $r(k)$ – autokoreliacinės funkcijos koeficientai, $a(1)$, $a(2)$, …, $a(p)$ – TPM koeficientai.

Arba galima užrašyti matriciniu pavidalu:

$$\mathbf{R} = \mathbf{ar}. \tag{2.20}$$

Šią lygčių sistemą galima išspręsti, pavyzdžiui, atvirkštinės matricos metodu. Tačiau **R** yra simetrinė Tioplico (*Toeplitz*) matrica ir jai spręsti galima naudoti Levinsono – Durbino algoritmą (Makhoul 1975; Markel, Gray 1976), reikalaujantį mažiau skaičiavimo operacijų. Tam tikslui naudojamos rekurentinės išraiškos:

$$E_0 = r(0), \tag{2.21}$$

$$a_i^i = k_i = \{r(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} r(|i-j|)\} / E^{(i-1)}, i = 1, 2, ..., p, \tag{2.22}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}, \tag{2.23}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)}, \tag{2.24}$$

čia $r(i)$ yra autokoreliacinės funkcijos reikšmės, E – paklaidos signalo energija.

Skaičiavimo metu taip pat gaunami atspindžio arba dalinės koreliacijos (*PARCOR*) koeficientai k_i, $i = 1,2,...,p$, kurie skirti sistemos stabilumo patikrinimui (jų absoliutinio dydžio vertės turi būti mažesnės už 1).

Jei apskaičiuosime (2.16) išraiškos z transformaciją, gausime sistemos funkciją, kuri, kalbos signalo atveju, aproksimuoja balso trakto perdavimo funkciją:

$$H(z) = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}}. \tag{2.25}$$

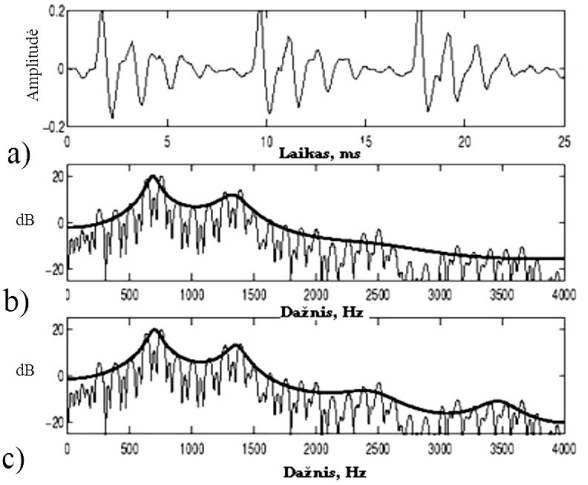
Apskaičiavę šią z transformaciją ant vienetinio apskritimo, t. y. kuomet $z=e^{j\omega}$, gausime balso trakto perdavimo funkcijos dažninę reakciją:

$$H(\omega) = \frac{G}{1 - \sum_{i=1}^{p} a_i e^{-j\omega i}}. \tag{2.26}$$

Iš šios dažninės reakcijos galime apskaičiuoti TPM spektrą, kuris aproksimuoja signalo spektro gaubtinę. Pavyzdžiui, (2.26) formulėje atlikę pertvarkymus, balso trakto dažninę amplitudės charakteristiką (amplitudinį TPM spektrą) galime išreikšti:

$$\left|H(\omega)\right| = \frac{G}{\sqrt{\left(1 + \sum_{i=1}^{p} a_i \cos(\omega i)\right)^2 + \left(\sum_{i=1}^{p} a_i \sin(\omega i)\right)^2}}. \tag{2.27}$$

Prognozavimo eilė – svarbus parametras sprendžiant atpažinimo uždavinius. 2.8 paveiksle pateikta kalbos signalo fragmentas, jo spektras ir spektro gaubtinės (TPM spektrai), gautos panaudojant skirtingas prognozavimo eiles *p*. Iš šio paveikslo matyti, kad panaudojant žemą prognozės eilę (p=6), gaunama apytiksli gaubtinė, turinti pagrindinę informaciją apie kalbą. Informacija apie kalbantįjį gali būti prarasta, bet lingvistinė informacija išsaugojama. Kai prognozės eilė aukštesnė (p=12), spektro gaubtinė atvaizduojama tiksliau, taip išsaugojama tiek lingvistinė, tiek ir kalbančiajam asmeniui būdinga informacija. Taigi, galima būtų daryti išvadą, kad žemesnė tiesinės prognozės eilė labiau tinka sprendžiant kalbos atpažinimo, o aukštesnė – kalbančiojo atpažinimo uždavinius.



2.8 pav. TPM analizės palyginimas panaudojant skirtingas prognozės *p* eiles (Mary *et al.* 2004): a) vokalizuotas kalbos signalo fragmentas; b) šio signalo spektras ir jo gaubtinė, kai *p*=6; c) šio signalo spektras ir jo gaubtinė, kai *p*=12

Fig. 2.8. Comparison of LPC analysis with different prediction order *p* (Mary *et al.* 2004): a) speech signal fragment; b) spectrum and its envelope when *p*=6; c) spectrum and envelope when *p*=12

2.3.5. Diskrečioji Furjė transformacija

Tolydinio ir diskretinio laiko signalai yra skirstomi į periodinius ir aperiodinius. Tolydinių periodinių signalų analizei yra skirtos Furjė eilutės, aperiodinių – Furjė transformacija. Tolydinio periodinio signalo spektras yra linijinis, atstumas tarp vienodai nutolusių linijų lygus pagrindiniam dažniui. Aperiodinio signalo spektrą gausime įsivaizduodami, kad jis yra periodinis, su periodu $T_p \to \infty$. Aperiodinio signalo atvirkštinės Furjė transformacijos formulė (sintezės lygtis):

$$s(t) = \int_{-\infty}^{\infty} X(F) e^{j2\pi Ft} dF \,. \tag{2.28}$$

Tiesioginė transformacija (analizės lygtis):

$$X(F) = \int_{-\infty}^{\infty} s(t) e^{-j2\pi Ft} dt \,. \tag{2.29}$$

Kaip matome, aperiodinio signalo spektras yra tolydinis.

Tolydinio laiko periodinių signalų dažnių diapazonas yra nuo -∞ iki +∞, dėl to yra signalų, kurie turi begalinį dažnių komponenčių skaičių. Diskretinio laiko signalų dažnių diapazonas yra vienintelis intervale (-π; π), todėl diskretinio laiko periodinis signalas su pagrindiniu periodu N gali būti sudarytas iš komponenčių, atskirtų $2\pi/N$ radianų. Todėl diskretinio laiko signalas daugiausiai gali turėti N komponenčių. Diskretinio laiko baigtinės energijos aperiodinių signalų dažninei analizei yra skirta Furjė transformacija:

$$X(\omega) = \sum_{n=-\infty}^{\infty} s(n) e^{-j\omega n} . \tag{2.30}$$

$X(\omega)$ yra periodinė su periodu 2π. Atvirkštinė Furjė transformacija (sintezės lygtis):

$$s(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\omega) e^{j\omega n} d\omega . \tag{2.31}$$

Baigtinės trukmės ilgio L seka turi Furjė transformaciją:

$$X(\omega) = \sum_{n=-\infty}^{L-1} s(n) e^{-j\omega n},\ 0 \le \omega \le 2\pi . \tag{2.32}$$

Tai reiškia, kad signalas $s(n) = 0$ už diapazono $0 \le n \le L-1$. Kai $X(\omega)$ imtys yra su vienodai nutolusiais dažniais $\omega_k = 2\pi k / N$, k=0, 1,..., N–1 ir $N \ge L$, gautos imtys yra:

$$X(k) = X\left(\frac{2\pi k}{N}\right) = \sum_{n=0}^{N-1} s(n) e^{-j2\pi kn/N} . \tag{2.33}$$

Ši formulė transformuoja seką $s(n)$ į $X(k)$. Kadangi dažnio imtys $X(k)$ yra gautos apskaičiuojant Furjė transformaciją $X(\omega)$ N vienodai nutolusių diskretinių dažnių aibėje, tai ši formulė vadinama $s(n)$ **diskrečiaja Furjė transformacija** (DFT). **Atvirkštinė diskrečioji Furjė transformacija** (ADFT):

$$s(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j2\pi kn/N},\ n = 0,1,...,N-1 . \tag{2.34}$$

Praktikoje dažniausiai naudojama **Greitoji Furjė transformacija** GFT (angl. – *Fast Fourier transform*), pasiūlyta septintame dešimtmetyje J. W. Cooley ir J. W. Tukey (Cooley, Tukey 1965). Pažvelgus į (2.33) išraišką matome, kad šiuo atveju algoritmo sudėtingumas yra O(N^2), kur N – nagrinėjamos signalo atkarpos ilgis. Tačiau DFT gali būti apskaičiuota greitesniu

būdu. Vienas iš pagrindinių reikalavimų yra tas, kad duomenų ilgis turi būti 2^M. Praktikoje dažnai daroma taip, jeigu signalo ilgis kitoks, jis papildomas nuliais iki artimiausio 2^M ilgio ir tuomet vykdomas GFT algoritmas. Nulius galima pridėti tiek nuo signalo pradžios, tiek nuo galo, ir tai nekeičia DFT rezultato. Pvz., jeigu signalo ilgis atskaitomis yra 224, jis papildomas nuliais iki 2^8= 256 ir vykdoma GFT. Šiuo atveju algoritmo sudėtingumas O($N \log_2 N$).

2.3.6. Diskrečioji kosinusų transformacija

Diskrečioji kosinusų transformacija (angl. – *discrete cosine transform*) (Huang *et al.* 2001) labai plačiai naudojama kalbos signalų apdorojime. Vienas iš jos privalumų yra tas, kad jos koeficientai labiau koncentruoti ties žemesniais indeksais, nei DFT koeficientai. Dėl to signalą galima aproksimuoti panaudojant mažiau koeficientų (Oppenheim *et al.* 1999). Yra keletas DKT apibrėžimų. Dažniausiai naudojama DKT išraiška:

$$C(k) = \sum_{n=0}^{N-1} s(n) \cos\left(\pi k \frac{n+1}{2N} \right), \; k = 0,1,...,N-1, \tag{2.35}$$

čia $s(n)$ – realus signalas. Atvirkštinė transformacija:

$$s(n) = \frac{1}{N}\left[C(0) + 2\sum_{k=1}^{N-1} C(k) \cos\left(\pi k \frac{n+1}{2N} \right) \right], \; n = 0,1,...,N-1. \tag{2.36}$$

Diskrečioji kosinusų transformacija naudojama skaičiuojant melų ir barkų skalės kepstro koeficientus ir t. t.

2.3.7. Melų skalės kepstro koeficientai

Kaip žinoma iš anatomijos, žmogaus vidinės ausies sraigė veikia kaip tam tikras spektro analizatorius (Huang *et al.* 2001). Taip pat gerai žinoma, kad žmogus dažnių skalę suvokia ne tiesiniu masteliu. Tam, kad surasti skales, atitinkančias žmogaus garso suvokimo sistemą, buvo atlikti įvairūs psichoakustiniai eksperimentai. Buvo nustatyta, kad yra tam tikros žmogaus suvokimo kritinės dažnių juostos (Fletcher 1940). Viena iš šių kritinių juostų skalių pavadinta barkų skale. Barkų skalė yra nuo 1 iki 24 barkų, atitinkanti 24 kritines girdėjimo juostas. Kita skalė yra melų skalė (Fletcher 1940), kuri iki 1 kHz yra beveik tiesinė, o toliau logaritminė. Melų skalė buvo išvesta atliekant eksperimentus su paprastais sinusoidiniais signalais. Melų skalė gali būti aproksimuota:

$$B(f) = 1125 \cdot \ln\left(1 + \frac{f}{700}\right). \tag{2.37}$$

Atvirkštinė transformacija (iš melų į hercus):

$$B^{-1}(b) = 700 \cdot \left(e^{\frac{b}{1125}} - 1\right). \tag{2.38}$$

Melų skalės kepstro koeficientai (MSKK) yra apibrėžti kaip realus kepstras, gaunami iš signalo kadro, padauginto iš lango funkcijos ir apskaičiuotos Furjė transformacijos. Jie skiriasi nuo realaus kepstro tuo, kad yra apskaičiuoti panaudojant specialius trikampius filtrus, suformuotus pagal netiesinę dažnių skalę. Apibrėžiame filtrų rinkinį, susidedantį iš I trikampių filtrų. i-tasis filtras apibrėžiamas:

$$H(i,f) = \begin{cases} 0, & \text{kai } f < f(i-1); \\ 2\dfrac{f - f(i-1)}{(f(i+1) - f(i-1))(f(i) - f(i-1))}, & \text{kai } f(i-1) \le f \le f(i); \\ 2\dfrac{f(i+1) - k}{(f(i+1) - f(i-1))(f(i+1) - f(i))}, & \text{kai } f(i) \le f \le f(i+1); \\ 0, & \text{kai } f > f(i+1). \end{cases} \tag{2.39}$$

Šie filtrai skaičiuoja vidurkinį spektrą aplink kiekvieną centrinį dažnį su vis platėjančiomis dažnių juostomis. Ribiniai dažniai $f(i)$ išsidėstę pagal melų skalę. Juos galima išreikšti:

$$f(i) = \left(\frac{N}{F_{\mathrm{d}}}\right) B^{-1}\left(B(F_{\mathrm{a}}) + i \cdot \frac{B(F_{\mathrm{v}}) - B(F_{\mathrm{a}})}{I+1}\right), \tag{2.40}$$

čia N – GFT dydis, I – filtrų skaičius, F_{d} – diskretizacijos dažnis (hercais), F_{a} ir F_{v} – žemiausias ir aukščiausias filtro dažnis (hercais). B^{-1} – duotas formulėje (2.38).

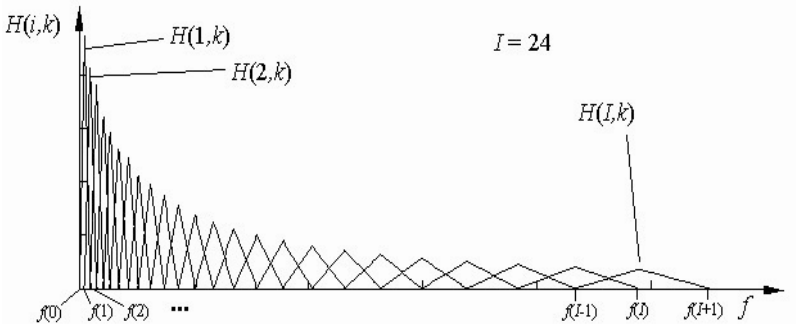
Energijos logaritmą filtro išėjime (melų skalės spektro koeficientus – MSSK) galime išreikšti:

$$C(i) = \ln\left(\sum_{k=0}^{N-1} H(i,k) \cdot |S(k)|^2\right), \; i = 1,2,...,I, \tag{2.41}$$

čia $S(k)$ – signalo Furjė transformacijos koeficientų (signalo spektro) amplitudės.

Melų skalės kepstrą galime išreikšti pasinaudoję diskrečiaja kosinusų transformacija:

$$c(j) = \sum_{i=0}^{I-1} C(i) \cdot \cos\left(\pi n \frac{i-1}{2I} \right), \quad j = 0,1,..., I\,. \tag{2.42}$$



2.9 pav. Trikampių filtrų rinkinys, naudojamas MSKK skaičiavimui, kai I=24

Fig. 2.9. Filterbank of triangular filters for MFCC calculation, when I=24

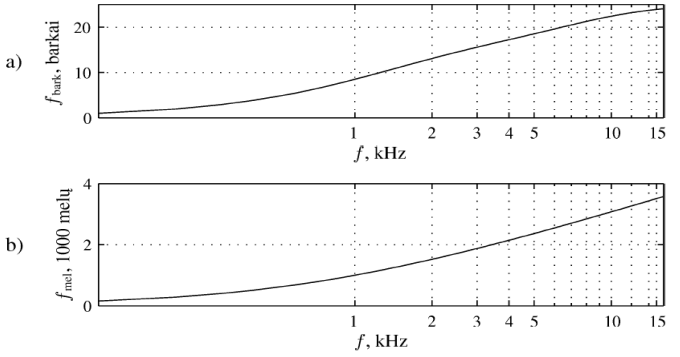
Melų skalės kepstro koeficientų panaudojimo atveju filtrų skaičius I parenkamas priklausomai nuo keliamos užduoties ir paprastai kinta nuo 24 iki 40 (2.9 paveikslas). Kalbos atpažinime jis paprastai imamas lygus 13.

2.3.8. Barkų skalės kepstro koeficientai

Barkų skalės kepstro koeficientai apskaičiuojami lygiai taip pat kaip ir melų skalės kepstro koeficientai. Skiriasi tik aproksimacijos formulė. Barkų skalė gali būti aproksimuota:

$$\text{Bark}(f) = 13\arctan\left(\frac{0,76f}{1000} \right) + 3,5\arctan\left(\frac{f}{7500} \right)^2. \tag{2.43}$$

Toliau, lygiai taip pat kaip ir skaičiuojant MSKK, sudaromi trikampiai filtrai, apskaičiuojamas energijos logaritmas kiekvieno filtro išėjime bei, taikant diskrečiąją kosinusų transformaciją, apskaičiuojami barkų skalės kepstro koeficientai. Hercų transformacija į barkus ir melus parodyta 2.10 paveiksle.



2.10 pav. Dažnių skalės (Tamulevičius 2008): a) barkų; b) melų

Fig. 2.10. Frequency scales (Tamulevičius 2008): a) bark; b) mel

2.3.9. Tiesinės prognozės modelio kepstras

Tai, ko gero, vieni iš labiausiai paplitusių ir duodančių geriausius atpažinimo rezultatus iš išvestinių tiesinės prognozės modelio parametrų. Tiesinės prognozės kepstro koeficientai (TPMK) gali būti gauti iš tiesinės prognozės modelio (TPM) spektro logaritmo, pritaikius diskrečiąją kosinusų transformaciją. Tačiau šiems kepstro koeficientams skaičiuoti yra ir kitų būdų, tam dažniausiai naudojamos rekurentinės išraiškos (Rabiner, Schafer 1978):

$$c_0 = \ln b^2 , \tag{2.44}$$

$$c_m = a_m + \sum_{k=1}^{m-1} (\frac{k}{m}) c_k a_{m-k}, \ m = 1,2,...,p , \tag{2.45}$$

$$c_m = \sum_{k=1}^{m-1} (\frac{k}{m}) c_k a_{m-k}, \ m > p , \tag{2.46}$$

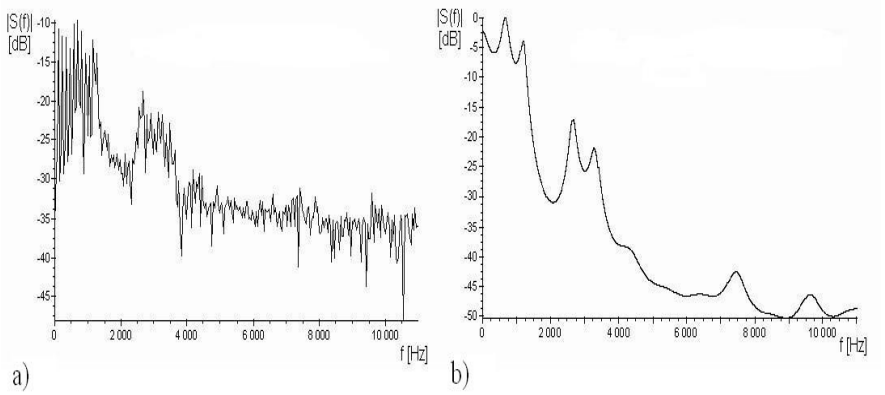
čia p – TPM modelio eilė, m – kepstro koeficientų eilė.

Kaip matome, kepstro koeficientų eilė gali būti didesnė už TPM modelio eilę. Praktikoje naudojama m=12,…, 22.

2.3.10. Foneminiai kalbos signalų požymiai

Formantinė kalbos struktūra. Yra teigiama, kad kalba susideda iš tam tikrų garsų, kurie gaunami žadinimo signalui sąveikaujant su kitais kalbos

padargais. Tariant atskirus garsus, pvz. „A“ ir „L“ skiriasi balso trakto konfigūracija bei žadinimo signalas. Vokalizuotų garsų energija yra stipriausia. Paėmę vokalizuoto garso fragmentą, pvz. fonemą „A“ ir atlikę jo Furjė transformaciją, gausime šios fonemos signalo spektrą. Kaip matome 2.11 paveiksle a) šiame spektre galima įžvelgti tam tikrus maksimumus, kurie vadinami *formantėmis*, bei minimumus, vadinamus *antiformantėmis*. Teigiama, kad dažnių juostoje 200–5 000 Hz galima išskirti 3–5 formantes. Kiekvieną ištartą garsą atitinka tam tikra balso trakto konfigūracija, formančių padėtis dažnių srityje bei amplitudės dydis. Formančių padėtis priklauso nuo tariamo garso bei asmens anatominių savybių. Nustatyta, kad asmuo sąmoningai gali keisti pirmų dviejų formančių padėtį, tuo tarpu aukštesnių formančių padėtis priklauso nuo asmens anatominių ypatumų.



2.11 pav. a) signalo kadro Furjė transformacija (spektras); b) perdavimo funkcija, apskaičiuota iš TPM parametrų (TPM spektras)

Fig. 2.11. Fourier transform (spectrum) of the signal frame a); transfer function calculated from the LPC parameters (LPC spectrum) b)

Formančių išsidėstymas geriausiai matomas nagrinėjant balso trakto perdavimo funkciją arba balso trakto spektrus (2.11 paveikslas b)). Šie spektrai gana plačiai naudojami tiek kalbos, tiek ir asmens atpažinime (Zilovic *et al.* 1998).

Formančių suradimo metodai. Formantės (Rabiner, Schafer 1978) susidaro sklindant garso bangai balso traktu. Garso banga iš dalies atsispindi nuo lūpų, grįždama gali interferuoti su balso traktu sklindančia banga ir sudaro akustinį rezonansą, kuris apsprendžia akustinį maksimumą. Vidutinį atstumą tarp

formančių apsprendžia vidutinis balso trakto ilgis L. i-tosios formantės dažnį galima išreikšti:

$$F(i) = \frac{C}{2L} \cdot \left(i - \frac{1}{2} \right), \tag{2.47}$$

čia C – garso bangos sklidimo greitis, L – vidutinis balso trakto ilgis.

Formantė tai yra dažninis rezonansas, kuris gali būti charakterizuotas ne tik dažniu, bet ir amplitude, kuri priklauso nuo rezonanso stiprumo. Formantės plotis tai rezonansinio piko plotis ties ta vieta, kur amplitudė pasiekia ½ maksimumo. Kalbant keičiasi balso trakto konfigūracija, dėl to kinta formančių padėtis, tačiau jų skaičius lieka pastovus. Tariant vokalizuotus garsus balso traktas kinta pakankamai lėtai, nevokalizuotų garsų trukmė yra žymiai mažesnė, formantiniai maksimumai išnyksta arba jų skaičius pasikeičia.

Formančių nustatymas nėra paprastas uždavinys. Taip yra todėl, kad esant tam tikriems balso trakto parametrams kai kurie spektro maksimumai spektro gaubiamojoje išnyksta, todėl juos apskaičiuoti iš spektro gaubiamosios yra sudėtinga, o kartais ir neįmanoma.

Tam tikslui yra alternatyvus būdas, vadinamas *spektrinių porų* (Kabal, Ramachandran 1986) metodu. Tarkime, turime signalo kadrą ir jame apskaičiuotus tiesinės prognozės modelio koeficientus a_i, i=1,…, p. Šie koeficientai aproksimuoja balso trakto perdavimo funkciją (2.25). Šios funkcijos polinomas:

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} \,. \tag{2.48}$$

Šio polinomo veidrodinis atspindys:

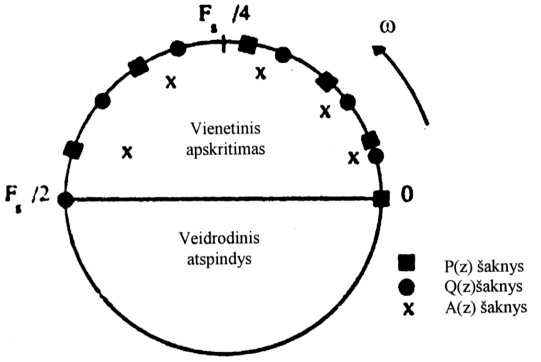
$$B(z) = z^{-(p+1)} A(z^{-1}) \,. \tag{2.49}$$

Sudėję ir atėmę šiuos polinomus gauname suminį ir skirtuminį polinomus:

$$P(z) = A(z) + B(z) \,; \tag{2.50}$$

$$Q(z) = A(z) - B(z) \,. \tag{2.51}$$

Šių polinomų šaknys išsidėstę ant vienetinio apskritimo ir sudaro spektrines poras (2.12 paveikslas).



2.12 pav. Polinomų $P(z)$, $Q(z)$, $A(z)$ šaknų išsidėstymas ant vienetinio apskritimo

Fig. 2.12. Roots of polynomials $P(z)$, $Q(z)$, $A(z)$ on unit cicle

Matome, kad polinomo $A(z)$ šaknys yra tarp polinomų $P(z)$ ir $Q(z)$ šaknų. Pažymėję polinomo $A(z)$ šaknis $z_1, z_1^*, z_2, z_2^*, ..., z_n, z_n^*$, čia z^* – kompleksiškai jungtinės šaknys:

$$A(k) = \left| \frac{(1 - z(k))(1 - z^*(k))}{(z - z(k))(z - z^*(k))} \right|^2, \tag{2.52}$$

čia $z=\exp(2\pi jF(k)T)$, $z(k)=|z(k)|\exp(2\pi jF(k)T)$, T – diskretizacijos periodas.

Iš čia gauname, kad formantės dažnis:

$$F(k) = \frac{1}{2\pi T} \operatorname{Im}(\ln(z(k))), \tag{2.53}$$

o formantės plotis:

$$B(k) = \frac{1}{2\pi T} \operatorname{Re}\left(\frac{1}{\ln(z(k))} \right). \tag{2.54}$$

Tiesiogiai rasti formantes iš šių formulių dažnai neįmanoma. Tai padaryti ne visada įmanoma ir iš TPM spektro gaubiamosios. Dėl to, norint apytiksliai apskaičiuoti formantes, galima apskaičiuoti spektrines poras ir vieną iš spektrinės poros dažnių priskirti atitinkamai formantei (Weber *et al.* 2002).

Formančių amplitudės skaičiuojamos iš TPM koeficientų (Rabiner, Schafer 1978):

$$S(\omega_k) = \frac{G}{\sum_{s=1}^{M+1} B(s)\cos\left(\frac{\omega_k s}{T}\right)}, \tag{2.55}$$

čia G – paklaidos signalo energija, priimama G=1000, T – diskretizacijos periodas, ω_k=2πF_k, F_k – k-tos formantės dažnis,

$$B(1) = \sum_{i=1}^{M+1} a_i , \tag{2.56}$$

$$B(s) = 2\sum_{i=0}^{M+1-s} a_i a_{i+s} , \tag{2.57}$$

čia s=2,…, p+1, $B(s)$ – TPM parametrų s-tos eilės autokoreliacija.

2.3.11. Vilnelių transformacijos požymiai

Dar 20 metų atgal pagrindine signalų analizės priemone buvo Furjė transformacija ir Furjė eilutės. Tai yra signalo pavaizdavimas harmoninių virpesių (sinusoidžių) suma su tam tikromis amplitudėmis. Toks pavaizdavimas atskleidžia visą signalo dažnių turinį, t. y. Furjė spektras parodo, kokio dažnio komponentės sudaro signalą. Kadangi harmoninės funkcijos nėra lokalizuotos laiko srityje, toks pavaizdavimas atskleidžia tik globaliąsias signalų savybes, t. y. iš jų negalima sužinoti tikslios signalo trūkio vietos ir pan.

Vilnelių transformacija (Chui 1992; Daubechies 1992; Jawerth, Sweldens 1994; Mallat 1998), panašiai kaip ir Furjė transformacija, atvaizduoja signalą tam tikrų funkcijų suma, tik vietoje harmonikų $\{e^{j\omega n}\}$ naudojama funkcijų sistema $\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\Psi\left(\frac{b-t}{a}\right)$. T. y. visos šios funkcijos gaunamos iš vienos fiksuotos funkcijos, ją perstumiant bei keičiant jos mastelį. Funkcija Ψ(t) vadinama vilnele, jei:

- Ψ(t) yra tolydi;
- Ψ(t) yra integruojama per visą laiko ašį;
- $\int_{-\infty}^{\infty} \Psi(t)dt = 0$.

Tolydinė vilnelių transformacija gali būti užrašyta:

$$Wf(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\Psi\left(\frac{b-t}{a}\right)dt , \tag{2.58}$$

čia a – mastelio parametras, b – postūmio parametras.

Taigi, priešingai nei Furjė transformacija, vilnelių transformacija apibrėžta nevienareikšmiškai, t. y. panaudojant skirtingas vilneles gaunamas skirtingas vilnelių spektras. Vilnelių parinkimas priklauso nuo keliamos užduoties.

Diskrečioji vilnelių transformacija. Šiuo atveju naudojamos diskrečiosios parametrų *a* ir *b* reikšmės, kurios užduodamos laipsninėmis funkcijomis:

$$a = a_0^{-m}, b = ka_0^{-m},\ m,k \in I, \tag{2.59}$$

kur *m* – mastelio parametras, *b* – postūmio parametras.

Tuomet:

$$\Psi_{mk}(t) = \mid a_0 \mid^{m/2} \Psi(a_0^m t - k). \tag{2.60}$$

Tiesioginė transformacija gali būti išreikšta:

$$c_{mk} = \int_{-\infty}^{\infty} s(t)\Psi_{mk}(t)dt. \tag{2.61}$$

Atvirkštinės transformacijos formulė (sintezės lygtis):

$$s(t) = \frac{1}{C_\Psi} \sum_{m=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} c_{mk}\Psi_{mk}(t). \tag{2.62}$$

Kalbos technologijose požymių išskyrimui vilnelės buvo panaudotos dvejopai: vietoje diskrečiosios kosinusų transformacijos (Tufekci, Gowdy 2000) ir tiesiogiai pritaikytos kalbos signalui. Antruoju atveju, požymiais buvo naudojami vilnelių koeficientai su didele energija (Long, Datta 1996). Taip pat buvo panaudotos vilnelių paketinės bazės (Farooq, Datta 2002; Sarikaya, Hansen 2000), panaudojant Daubechies ortogonalius filtrus su 32 ir 12 koeficientų, kurios sudalindavo signalo dažnių juostą artimai melų skalei (panašiai kaip trikampiai filtrai, skaičiuojant MSKK). Sarikaya (Sarikaya *et al.* 1998) panaudojo šiuos požymius kalbančiojo identifikavime ir paskelbė, kad jie atpažinimo tikslumu pralenkė MSKK. Vilnelių panaudojimas kalbančiojo atpažinime aptartas ir (Siafarikas *et al.* 2004).

2.3.12. Spektro dinamikos požymiai

Spektriniai kalbos signalų požymiai skaičiuojami trumpalaikiuose kalbos signalų intervaluose – kadruose. Tai yra lėtai besikeičiančio balso trakto momentinis pavaizdavimas pastoviais parametrais. Šiuose požymiuose nėra informacijos apie balso trakto kitimą laike.

Kalbant keičiasi balso trakto konfigūracija ir žadinimo signalas. Šie pokyčiai atsispindi požymių vektorių pasikeitime. Šių pokyčių greitis priklauso nuo kalbėjimo stiliaus, greičio, kalbos. Kai kurie iš dinaminių spektrinių požymių taip pat priklauso ir nuo kalbančiojo asmens.

delta- ir delta-delta požymiai

Vienas iš plačiausiai žinomų metodų, kaip gauti spektrinių požymių dinaminę informaciją yra delta- požymių skaičiavimas (Huang *et al.* 2001). Šie delta- požymiai yra prijungiami prie statinių požymių. Delta požymiai gaunami skaičiuojant statinių požymių laiko išvestines. Dažnai skaičiuojamos delta požymių laiko išvestinės ir gaunami, taip vadinami, delta-delta požymiai.

delta- ir delta-delta požymiai naudojami su tam tikrais požymiais – ypač su kepstru ir jo variantais (TPMK, MSKK ir t. t.) (Ariyaeeinia, Sivakumaran 1995; Hume 1997).

Yra du pagrindiniai metodai kaip vertinti išvestines (Hansen, Proakis 2000; Furui 1981; Hume 1997): diferencijuojant ir pritaikant polinominę kreivę. Tegul $f_k[i]$ yra k-tojo kadro i-tasis požymis. Diferencijuojant, k-tojo požymių vektoriaus f_k i-tosios komponentės delta parametrą galima išreikšti:

$$\Delta f_k[i] = f_{k+M}[i] - f_{k-M}[i], \tag{2.63}$$

čia M yra tipiškai 1, 2 ar 3 kadrai. Diferencijavimas atliekamas kiekvienai požymių vektoriaus komponentei i ir taip gaunamas delta- požymių vektorius. Diferencijavimo metodas yra paprastas, bet, kadangi jis veikia kaip aukštų dažnių filtravimo operacija požymių srityje, yra pakeliami triukšmai (Furui 1981). Dėl šios priežasties, panaudojant polinominės kreivės pritaikymą parametrui laiko srityje, galima gauti geresnius rezultatus.

RASTA apdorojimas

Kaip alternatyva dinaminių požymių radimui yra naudojamas RASTA apdorojimas (Hermansky 1994). RASTA yra pagrįstas žmogaus klausos mechanizmo modeliu. Žmogaus ausis yra jautresnė tam tikriems moduliacijos dažniams ir RASTA apdorojimas siekia išfiltruoti visus nereikšmingus moduliacijos dažnius. RASTA ir susiję metodai kalbančiojo atpažinime panaudoti (Reynolds 1994; Hardt, Fellbaum 1997).

2.4. Kalbos signalų segmentavimas

Kalbos signalų segmentavimas (angl. – *voice activity detection*) yra labai svarbus etapas kuriant kalbos ir kalbančiojo atpažinimo sistemas. Kalbos ir

kalbančiojo atpažinimo sistemose naudojamas „kalbos detektorius“, kurio paskirtis surasti kalbą atitinkančius signalo kadrus (atskirti juos nuo triukšmo) tolesniam apdorojimui. „Kalbos detektoriai“ skiriasi priklausomai nuo požymių, kuriuos jie naudoja ir nuo šių požymių klasifikavimo metodų.

Kalbos fragmentų išskyrimui dažniausiai naudojami šie požymiai (Zilca *et al.* 2004):

- Kadro energija.
- Nulių kirtimų dažnis.
- Tie patys požymiai, naudojami kalbos ar kalbančiojo atpažinimo sistemoje (MSKK ir pan.).
- Sudėtiniai požymiai. Dažnai naudojama: energija, pagrindinis tonas, vokalizavimo lygis, nulių kirtimų dažnis (angl. – *zero crossing rate*) ir t. t.

Dažniausiai naudojami šie klasifikavimo metodai (Zilca *et al.* 2004):

- Fiksuotas energijos slenkstis. Kadrai, kurių energija didesnė už slenkstį, klasifikuojami kaip kalbos signalas. Veikia realiame laike.
- Energijos slenkstis, priklausantis nuo pasakymo. Priklauso nuo iš ištarimo atitinkančio kalbos signalo gautos energijos statistikos. Neveikia realiame laike.
- Triukšmo lygio sekimo arba būsenos mechanizmas. Paprastai naudojama su energija ar sudėtiniais požymiais. Veikia artimai realiame laike.
- Pagrįsti modeliu. Požymių vektoriai, gauti iš MSKK, reikalauja sudėtingesnio palyginimo, panašaus į statistinio modeliavimo metodus, naudojamus kalbos/kalbančiojo atpažinimui. Tie patys metodai, naudojami kalbos/kalbančiojo atpažinimui, gali būti panaudoti ir kalbos išskyrimui. Tai taip pat gali veikti realiame laike.

Kalbos išrinkimas yra būtina kalbančiojo atpažinimo sistemos dalis. Ji reikalinga tam, kad kalbėtojo modeliai būtų apmokyti tik kalbos signalo požymiais, taip pat atpažinimui būtų pateikti tik kadrai, atitinkantys kalbą. Kalbančiojo asmens atpažinimo tikslumas priklauso nuo „kalbos detektoriaus“ ypatingai nepalankiose sąlygose.

Jei „kalbos detektorius“ įvairius foninius triukšmus priskirs kalbai, padidės „klaidingo priėmimo“ tikimybė, jei atpažinimui bus pateiktas kalbos signalas su panašiais foniniais triukšmais. Jei „kalbos detektorius“ atmes kalbos signalo kadrus, klaidos vėl didės, kadangi trūks kalbos duomenų.

Toliau labai trumpai aptarsime kelis nesudėtingus klasikinius metodus, taikomus kalbos signalų segmentavimui, tai *triukšmo slenksčio* ir *energijos slenksčio* bei *nulių kirtimo dažnio* metodus.

2.4.1. Triukšmo slenkstis

Vienas iš pačių paprasčiausių metodų yra triukšmo slenksčio nustatymas. Tuo tikslu imamas tam tikro ilgio triukšmo signalas (pvz. nuo garso įrašo pradžios imama keliasdešimt milisekundžių) ir skaičiuojamas triukšmo slenkstis:

$$Thr = \frac{1}{N_{\text{tr}}} \sum_{n=1}^{N_{\text{tr}}} \left| s_{\text{tr}}(n) \right|, \tag{2.64}$$

čia N_{tr} – triukšmo signalo ilgis (atskaitomis), $s_{\text{tr}}(n)$ – triukšmo signalas. Nors šis metodas yra labai paprastai realizuojamas, jį panaudojant, esant aukštesniam triukšmo lygiui, gaunami labai prasti rezultatai. Dėl to jį galima taikyti tik esant labai geroms įrašymo sąlygoms.

2.4.2. Energija ir nulių kirtimų dažnis

Kiti, plačiai naudojami, būdai atskirti kalbos signalui nuo triukšmo yra energija ir nulių kirtimų dažnis. Vokalizuotų garsų energija yra didelė, dėl to nustačius tam tikrą slenkstį, galima kiekvieno kadro energiją lyginti su šiuo slenksčiu. Tačiau šis metodas nelabai tiksliai nustato nevokalizuotus garsus, ypač esant aukštesniam triukšmo lygiui, kadangi nevokalizuotų garsų energija yra nedidelė.

Vokalizuotų garsų nulių kirtimų dažnis yra mažesnis už triukšmo ar nevokalizuotų garsų, todėl šis metodas taip pat gali būti panaudotas „kalbos detektoriuje“. Tačiau jis taip pat nelabai gerai atskiria nevokalizuotus garsus nuo triukšmo.

2.5. Kalbančiojo modeliavimo ir požymių vektorių palyginimo metodai

Šiame skyriuje bus aptarti kalbančiojo asmens modeliavimo metodai ir būdai, kaip rasti panašumo matą tarp kalbėtojo modelio ir atpažinimui skirtos frazės požymių vektorių. Šie abu metodai, bendrai apjungus, vadinami *klasifikacijos metodu. Kalbančiojo modeliavimas* tai yra procesas, kurio metu kalbančiojo modelis yra sukuriamas ir įtraukiamas į atpažinimo sistemą. Modelis sukuriamas turint to asmens mokymui skirtų frazių požymių vektorius. *Palyginimas* – tai yra procesas, kurio metu skaičiuojamas panašumo įvertis tarp kalbėtojo modelio ir nežinomo asmens ištartos frazės požymių vektorių (Campbell 1997).

Yra du pagrindiniai klasifikacijos problemos sprendimo būdai: etalonų palyginimas ir stochastinis palyginimas (Campbell 1997). Abu metodai gali būti tiek priklausomi nuo ištarto teksto, tiek ir nepriklausomi. Stochastinis metodas sukuria tikimybinį kalbančiojo modelį, kuris nurodo kintančias laike kalbos signalo charakteristikas (Naik 1990). Naudojant šį metodą kalbančiojo modelis aprašomas tikimybiniais požymių vektorių pasiskirstymais ir klasifikacija pagrįsta tikimybėmis arba tikėtinumais.

Iš nepriklausančių nuo ištarto teksto kalbančiojo atpažinimo metodų bene populiariausi yra Gauso mišinių modeliai (GMM) ir vektorinio kvantavimo (VK) metodas. Taip pat plačiai naudojami dirbtinių neuronų tinklai (DNT) ir ne taip seniai pradėtos taikyti atraminių vektorių mašinos (AVM). Priklausomame nuo ištarto teksto kalbančiojo atpažinime naudojami tie patys metodai kaip ir kalbos atpažinime. Populiariausi iš jų yra dinaminis laiko skalės kraipymas (DLK) (Rabiner 1978; Sakoe 1979; White, Neely 1976) ir paslėptieji Markovo modeliai (PMM) (Rabiner, Juang 1986). Toliau trumpai aptarsime nepriklausomus nuo ištarto teksto kalbančiojo atpažinimo metodus.

2.5.1. Gauso mišinių modeliai

Vienas iš pačių populiariausių kalbančiojo modeliavimo metodų yra Gauso mišinių modeliai (Reynolds, Rose 1995) (angl. – *Gaussian mixture models*). Šis metodas priklauso stochastiniam modeliavimui ir modeliuoja statistinį požymių kitimą. Tai yra statistinis pavaizdavimas kaip kalbėtojas taria garsus (Reynolds 2002).

Gauso mišinių tankis yra M komponenčių tankių pasverta suma, pavaizduota 2.13 paveiksle, kurią galima išreikšti:

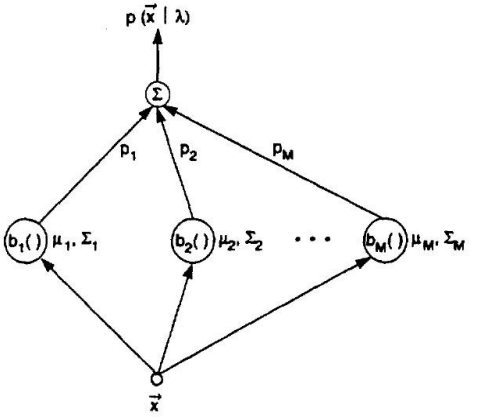
$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}), \tag{2.65}$$

čia $\vec{x}$ – D-dimensinis atsitiktinis vektorius, $b_i(\vec{x})$ – komponenčių tankiai, i=1, 2,..., M, p_i, i=1, 2,..., M – mišinių svorių koeficientai.

Kiekvienas komponentės tankis $b_i(\vec{x})$, i=1, 2,..., M yra Gauso funkcija, kurią galima išreikšti:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}\left|\sum_i\right|^{1/2}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)'\sum_i^{-1}(\vec{x}-\vec{\mu}_i)}, \tag{2.66}$$

čia $\vec{\mu}_i$ – vidurkio vektorius, $\sum_i$ – kovariacinė matrica.



2.13 pav. M komponenčių Gauso mišinių tankio pavaizdavimas (Reynolds, Rose 1995)

Fig. 2.13. Depiction of M component Gaussian mixture density (Reynolds, Rose 1995)

Svorių koeficientai turi tenkinti ribojimus:

$$\sum_{i=1}^{M} p_i = 1. \tag{2.67}$$

Gauso mišinių tankis yra parametrizuojamas jo komponenčių tankių vidurkių vektoriais, kovariacinėmis matricomis ir mišinių svoriais. Visi šie parametrai bendrai žymimi rinkiniu:

$$\lambda = \left\{ p_i, \vec{\mu}_i, \sum_i \right\} \quad i = 1,2,...,M \,. \tag{2.68}$$

Taigi, kiekvienas kalbėtojas yra atstovaujamas tam tikru GMM parametrų rinkiniu λ.

GMM gali turėti kelias skirtingas formas, priklausomai nuo kovariacinių matricų parinkimo. Kiekviena mišinio komponentė gali turėti savo kovariacinę matricą, viena kovariacinė matrica gali būti visoms Gauso komponentėms (vieno kalbėtojo modeliui) ir viena kovariacinė matrica visiems kalbėtojams. Taip pat kovariacinė matrica gali būti pilna arba simetrinė.

Yra dvi pagrindinės prielaidos (Reynolds, Rose 1995), leidžiančios panaudoti Gauso mišinių tankius, siekiant sukurti kalbančiojo modelį. Pirmoji yra tai, kad individualūs komponenčių tankiai gali modeliuoti tam tikrų akustinių klasių rinkinį. Galima kalbėtojo balsą laikyti tam tikra akustine erdve, kuri gali

būti sudaryta iš tam tikrų akustinių klasių, atvaizduojančių tam tikrus fonetinius įvykius, kaip balsius, nosinius garsus ar pučiamuosius priebalsius. Šios akustinės klasės atspindi tam tikras bendras balso trakto konfigūracijas, kurios priklauso nuo kalbančiojo ir gali būti panaudotos siekiant modeliuoti kalbančiojo tapatybę. Bet kurios *i*-tosios akustinės klasės spektro forma gali būti apibūdinta vidurkiu $\vec{\mu}_i$, o vidutinės spektro formos svyravimas gali būti nusakytas kovariacine matrica $\sum_i$.

Antroji Gauso mišinių tankių panaudojimo kalbančiojo atpažinimui prielaida yra tai, kad bazinių Gauso funkcijų tiesinė kombinacija gali atvaizduoti didelę pavyzdžių pasiskirstymų klasę. Tarkime, klasikinis, vienmodalinis Gausinis kalbančiojo modelis, kalbančiojo požymių pasiskirstymą apibūdina vidurkio vektoriumi (pozicija) ir kovariacine matrica (elipsine forma). Vektorinio kvantavimo modelis kalbantįjį apibūdina tam tikru diskrečiuoju požymių rinkiniu. Šiuo atveju GMM yra kaip hibridinis modelis tarp šių dviejų, kadangi naudoja Gausinių funkcijų diskretųjį rinkinį, kur kiekviena iš šių funkcijų turi savo vidurkį ir kovariacinę matricą.

Maksimalus tikėtinumas ir parametrų vertinimas

Turint kalbančiojo ištartas mokymui skirtas frazes, kalbančiojo modelio mokymo tikslas yra įvertinti GMM parametrus λ, kurie geriausiai atitinka mokymo metu gautų požymių vektorių pasiskirstymą. Nors yra keletas GMM parametrų vertinimo būdų (McLachlan 1988), bet labiausiai paplitęs metodas yra maksimalaus tikėtinumo (ML) vertinimas.

Maksimalaus tikėtinumo vertinimo tikslas yra rasti modelio parametrus, kurie maksimizuos GMM modelio tikėtinumą. Turint *T* mokymo vektorių seką $X = \{\vec{x}_1,...,\vec{x}_T\}$, GMM tikėtinumas gali būti užrašytas:

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda) . \qquad (2.69)$$

Ši išraiška yra parametrų λ netiesinė funkcija, todėl tiesioginis maksimizavimas yra neįmanomas. Tačiau šis parametrų vertinimas gali būti atliktas iteraciniu būdu, panaudojant matematinės vilties maksimizavimo algoritmo (MVM) atskirą atvejį (Dempster *et al.* 1977). Šio algoritmo pagrindinė idėja, pradžioje turint pradinį modelį λ, įvertinti naują modelį $\overline{\lambda}$, kad $p(X \mid \overline{\lambda}) \geq p(X \mid \lambda)$. Tuomet naujas modelis vėl tampa pradiniu modeliu ir vykdoma sekanti iteracija. Tai tęsiama tol, kol pasiekiamas tam tikras konvergencijos slenkstis. Panašiai atliekama ir vertinant PMM parametrus, panaudojant Baum – Welch pakartotinio vertinimo algoritmą (Baum *et al.* 1970).

Kiekvienos MVM iteracijos metu naudojamos pakartotinio vertinimo formulės, garantuojančios modelio tikėtinumo vertės monotonišką augimą.

Svorių koeficientai perskaičiuojami:

$$\overline{p}_i = \frac{1}{T}\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda). \tag{2.70}$$

Vidurkių perskaičiavimo formulė:

$$\vec{\overline{\mu}}_i = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)\vec{x}_t}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)}. \tag{2.71}$$

Dispersija perskaičiuojama:

$$\overline{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)(\vec{x}_t - \vec{\mu}_i)^2}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)}. \tag{2.72}$$

Aposteriorinė *i*-tos klasės tikimybė gali būti išreikšta:

$$p(i \mid \vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)}. \tag{2.73}$$

Dvi problemos (Reynolds, Rose 1995), su kuriomis susiduriama naudojant Gauso mišinių kalbančiojo modelius yra mišinių eilės M parinkimas ir pradinis modelių parametrų inicializavimas, prieš panaudojant matematinės vilties maksimizavimo algoritmą. Tačiau nėra gero teorinio pagrindimo šių problemų sprendimui, todėl tai dažniausiai nustatoma eksperimentiniu keliu.

Kalbančiojo identifikavimas

Kalbančiojo identifikavimo metu turime S kalbėtojų $S=\{1, 2,\ldots, S\}$, kurie atstovaujami atitinkamais GMM $\lambda_1, \lambda_2,..., \lambda_S$. Tikslas yra rasti kalbančiojo modelį, turintį stebimos požymių sekos maksimalią aposteriorinę tikimybę:

$$\hat{S} = \arg\max_{1 \le k \le S} \Pr(\lambda_k \mid X) = \arg\max_{1 \le k \le S} \frac{p(X \mid \lambda_k)\Pr(\lambda_k)}{p(X)}. \tag{2.74}$$

Antroji išraiška yra Bejeso formulė. Kadangi kalbėtojai vienodai tikėtini, gauname, kad:

$$\Pr(\lambda_k) = \frac{1}{S}. \tag{2.75}$$

Be to, $p(X)$ yra ta pati visiems kalbėtojams, gauname supaprastintą klasifikavimo formulę:

$$\hat{S} = \arg\max_{1 \le k \le S} \Pr(X \mid \lambda_k). \tag{2.76}$$

Skaičiavimams dažnai naudojami logaritmai. Be to, kadangi stebėjimų sekos yra nepriklausomos, kalbančiojo identifikavimo sistema skaičiuoja:

$$\hat{S} = \arg\max_{1 \le k \le S} \sum_{t=1}^{T} \log p(\vec{x}_t \mid \lambda_k), \tag{2.77}$$

kur $p(\vec{x}_t \mid \lambda_k)$ duota (2.69) formulėje.

2.5.2. Vektorinis kvantavimas

Vektorinio kvantavimo metodas (Juang *et al.* 1987) taip pat labai plačiai naudojamas nepriklausomame nuo ištarto teksto kalbančiojo atpažinime. Vienas iš pagrindinių jo privalumų – yra palyginti paprasta algoritmo realizacija, greitas kalbančiųjų asmenų frazių palyginimas, taip pat paprastesni reikalavimai mokymo duomenims (jų reikia mažiau) nei naudojant pvz. GMM ar PMM. Vienas iš didesnių trūkumų, – mažesnis atpažinimo tikslumas. Vektorinio kvantavimo idėja pagrįsta tuo, kad siekiant pagreitinti skaičiavimus, reikalingus dviems frazėms palyginti, visa kalbos signalą atitinkančių požymių vektorių aibė atvaizduojama mažesniu, iš anksto pasirinktu, požymių vektorių skaičiumi. Tam tikslui pradinė požymių vektorių erdvė padalinama į pasirinktą skaičių nepersikertančių sričių (klasterių), ir kiekvienas iš klasterių atvaizduojamas vidurkiniu vektoriumi (centroidu). Centroidų aibė sudaro taip vadinamą *kodinę knygą*, kuri ir yra kalbėtojo modelis. Centroidų skaičius (kodinės knygos dydis) yra žymiai mažesnis už pradinių požymių vektorių skaičių. Taigi, kodinė knyga efektyviai sumažina duomenų kiekį, išlaikydama svarbiausią informaciją apie tų duomenų pasiskirstymą (Gersho, Gray 1991).

Kodinių knygų kūrimui yra dvi metodų klasės: neprižiūrėti ir prižiūrėti mokymo algoritmai. Neprižiūrėtuose metoduose kiekvieno kalbėtojo kodinės knygos kuriamos atskirai vienos nuo kitos, tuo tarpu prižiūrėtuose metoduose yra taip nustatomos tarpusavio koreliacijos tarp kodinių knygų, kad jos turėtų kuo mažesnį persidengimą. Vienas iš pasiūlytų prižiūrėtų metodų: grupinis vektorinis

kvantavimas (GVQ – angl. *Group vector quantization*) (He *et al.* 1999). Remiantis šiuo metodu, kiekviena iš kodinių knygų kuriama atskirai ir po to jos keičiamos taip, kad gautųsi kuo didesni skirtumai tarp kalbėtojų.

Kodinių knygų kūrimui naudojami šie algoritmai (Kinnunen *et al.* 2000):

- Apibendrintas Lloydo algoritmas ALA (angl. *GLA – Generalized Lloyd algorithm*).
- Save organizuojantys žemėlapiai SOM (angl. – *Self-organizing maps*).
- Poriškai artimiausias kaimynas PNN (angl. – *Pair-wise nearest neighbor*).
- Pasikartojanti dalinimo technika, SPLIT.
- Atsitiktinė vietinė paieška RLS (angl. – *Randomized local search*).

Vienas iš svarbesnių dalykų vektoriniame kvantavime yra kodinės knygos dydžio parinkimas. Didinant kodinę knygą gaunami mažesni iškraipymai, tačiau tuomet didėja skaičiavimo operacijų skaičius. Remiantis eksperimentiniais rezultatais (Kinnunen *et al.* 2000) kodinės knygos dydis parenkamas iki 64 centroidų (turėtų būti apytiksliai lygus skirtingų garsų skaičiui frazėje). Kvantavimo iškraipymai tarp pradinių požymių vektorių $X = \{\vec{x}_1,...,\vec{x}_T\}$ ir kodinės knygos $C = \{\vec{c}_1,...,\vec{c}_K\}$ dažniausiai skaičiuojami kaip kvadratinių atstumų suma tarp požymių vektorių ir juos atitinkančių centroidų. Kvantavimo klaida tarp vektoriaus $\vec{x}_i$ ir kodinės knygos *C*:

$$d_q(\vec{x}_i, C) = \min_{c_j \in C} d(\vec{x}_i, \vec{c}_j), \tag{2.78}$$

čia $d(\vec{x}_i, \vec{c}_j)$ žymi atstumą tarp požymių vektorių $\vec{x}_i$ ir $\vec{c}_j$. Dažniausiai naudojami atstumai yra absoliutinis, Euklidinis, pasvertas Euklidinis ir Mahalanobis (Campbell 1977; Ong *et al.* 1996). Absoliutinio atstumo formulė:

$$d_c(x, y) = \sum_{i=1}^{N} |x_i - y_i|. \tag{2.79}$$

Euklidinio atstumo formulė:

$$d_E(x, y) = (x - y)^T (x - y) = \sum_{i=1}^{N} (x_i - y_i)^2. \tag{2.80}$$

Pasverto Euklidinio atstumo formulė:

$$d_M(x, y) = (x - y)^T \cdot \mathbf{D}^{-1} \cdot (x - y), \tag{2.81}$$

čia *x* ir *y* yra daugiadimensiniai požymių vektoriai, **D** yra svorių matrica (Campbell 1977; Ong *et al.* 1996). Kai **D** yra kovariacinė matrica, pasvertas Euklidinis atstumas dar vadinamas Mahalanobis atstumu.

Vidutiniai kvantavimo iškraipymai gali būti išreikšti:

$$D_Q(X,C) = \frac{1}{T}\sum_{i=1}^{T} d_q(\vec{x}_i, C). \tag{2.82}$$

Toliau aptarsime ALA arba dar vadinamą LBG (Linde *et al.* 1980) vektorinio kvantavimo algoritmą išsamiau. Tarkime, kad turime pradinę požymių vektorių erdvę $X = \{\vec{x}_1, ..., \vec{x}_T\}$, kur *T* yra požymių vektorių, gautų iš stacionarių kalbos intervalų, skaičius.

LBG algoritmą galime suskaidyti į šiuos etapus:

1. Skaičiuojamas nulinis centroidas.
2. Centroidas dalijamas į du, „iškraipant“ jį pagal formulę (čia pasirinkta reikšmė μ= 0,01):

$$c_n^+ = c_n(1+\mu), \tag{2.83}$$

$$c_n^- = c_n(1-\mu). \tag{2.84}$$

3. Artimiausio kaimyno paieška. Kiekvienam požymių vektoriui skaičiuojamas atstumas iki kiekvieno centroido. Kiekvienam požymių vektoriui randamas artimiausias centroidas ir šis vektorius priskiriamas artimiausiam centroidui.
4. Centroidų patikslinimas. Centroidas perskaičiuojamas pagal jam priskirtus požymių vektorius.
5. Kartojami 3 ir 4 žingsniai, kol patikslinus centroidą iškraipymų kitimas pasidaro mažesnis už pasirinktą.
6. Randamas centroidas, turintis didžiausius iškraipymus arba visiems centroidams kartojami žingsniai 2, 3, 4 ir 5, kol negauname norimo dydžio kodinės knygos, arba kodinės knygos iškraipymai pasidaro mažesni už pasirinktą slenkstį.

Nulinio centroido skaičiavimas. Visą pradinę požymių vektorių erdvę laikydami vienu klasteriu, galime apskaičiuoti šio klasterio svorio centrą – nulinį centroidą. Nulinis centroidas gaunamas kaip visą klasterį sudarančių požymių vektorių vidurkis.

Dviejų naujų centroidų sudarymas. Centroidas dalijamas į du jį „iškraipant“ pagal (2.83) ir (2.84) formules. Iškraipymas atliekamas skirtingai, priklausomai nuo naudojamų požymių vektorių. Pavyzdžiui, naudojant TPM

parametrus, „iškraipomi“ juos atitinkantys atspindžio (PARCOR) koeficientai (Lipeika, Lipeikienė 1995), tuo tarpu naudojant kepstrinius požymius, „iškraipomi“ patys kepstro koeficientai.

Požymių vektorių klasifikavimas. Po centroido dalijimo, atliekamas pradinės požymių vektorių aibės klasifikavimas. Klasifikuojama pagal artimiausio kaimyno taisyklę t. y. kiekvienas iš pradinių požymių vektorių $X = \{\vec{x}_1,...,\vec{x}_T\}$ priskiriamas tam centroidui, kurio atstumas iki šio požymių vektoriaus yra mažesnis. Atstumo skaičiavimas taip pat priklauso nuo pasirinktų požymių vektorių.

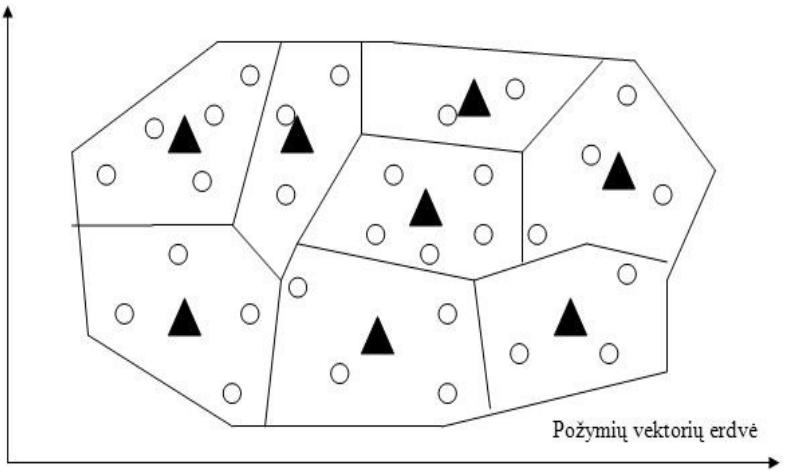
Taigi, atlikę parametrų klasifikaciją, gauname, kad pradinė požymių vektorių erdvė $X = \{\vec{x}_1,...,\vec{x}_T\}$ bus padalinta į du pradinius klasterius.

Centroidų padėties tikslinimas. Atlikus požymių vektorių klasifikaciją, toliau tikslinamos centroidų padėtys. Patikslintas centroidas gaunamas apskaičiavus jam priklausančio klasterio požymių vektorių vidurkį. Gauti naujų dviejų klasterių svorio centrai skirsis nuo prieš tai buvusių pradinių centroidų, gautų „iškraipius“ nulinį centroidą. Toliau vėl randame pradinių parametrų vektorių $X = \{\vec{x}_1,...,\vec{x}_T\}$ atstumus iki dviejų patikslintų centroidų, klasifikuojame parametrus pagal artimiausio kaimyno taisyklę (t. y. kiekvieną iš pradinių požymių vektorių $X = \{\vec{x}_1,...,\vec{x}_T\}$ priskiriame vienam ar kitam centroidui iki kurio ir tarp šio požymių vektoriaus bus mažesnis atstumas). Atlikus klasifikaciją, randami vidutiniai iškraipymai, gauti pradinę požymių vektorių aibę atvaizdavus dviem patikslintais centroidais. Jeigu vidutiniai iškraipymai sumažėja daugiau, nei pasirinktas slenkstis ε, toliau tęsiamas centroidų padėties tikslinimas, jei ne, tai ši procedūra nutraukiama.

Jeigu vidutiniai iškraipymai mažesni už pasirinktą slenkstį δ, tuomet klasterizacija nutraukiama. Priešingu atveju klasterizacija toliau tęsiama. Tai galima atlikti dviem būdais:

- Kiekvienas iš centroidų dalijamas į du (standartinis LBG metodas).
- Į du dalijamas vienas centroidas, turintis didžiausius iškraipymus (Lipeika, Lipeikienė 1995).

Pirmuoju atveju kiekvieno dalijimo metu klasterių skaičius didės du kartus (jis bus lygus 2^n, kur n – dalijimų skaičius). Antruoju atveju, kiekvieno naujo dalijimo metu, klasterių skaičius padidės vienetu. Galutinai atlikus klasterizaciją, gaunama kodinė knyga, kurią sudaro klasterius atitinkantys centroidai. Galutinai suklasterizuotos požymių vektorių erdvės, susidedančios iš aštuonių klasterių (centroidų) pavyzdys pateiktas 2.14 paveiksle.



2.14 pav. Galutinė požymių vektorių erdvės klasterizacija ir centroidų suradimas

Fig. 2.14. Final clusterization of feature space and calculation of the centroids

Kalbančiojo atpažinimo metu skaičiuojamas atitikimo įvertis (skirtumas) tarp atpažinimui skirtos frazės požymių vektorių ir visų kalbėtojų (identifikavimo atveju) saugomomis kodinėmis knygomis. Atitikimo įvertis dažnai skaičiuojamas taip pat kaip ir kvantavimo iškraipymai (2.82). Galima skaičiuoti ir vidutinę kvadratinę klaidą MSE (angl. – *Mean square error*):

$$MSE(X,C) = \frac{1}{T}\sum_{i=1}^{T}(d_q(\vec{x}_i,C))^2 . \qquad (2.85)$$

Asmuo, kuriam gaunama mažiausia klaida, gali būti grąžinamas kaip identifikavimo rezultatas.

2.5.3. Atraminių vektorių mašinos

Atraminių vektorių mašinos (angl. – *support vector machines*), vienas iš palyginti neseniai atsiradusių statistinių požymių klasifikavimo metodų, pasiūlytų Vapnick 1995 metais (Vapnick 1995), davusių labai gerus rezultatus vaizdų apdorojime, ypač veido atpažinime (Osuna *et al.* 1997). Šie klasifikatoriai veikia struktūrinės rizikos minimizavimo (Vapnick 1995) principu. Eksperimentai rodo, kad šie klasifikatoriai veikia ne prasčiau už kitus, tuo tarpu reikalauja mažiau mokymo duomenų. Pagrindinė atraminių vektorių mašinų (AVM) idėja yra dviejų skirtingų klasių mokymo duomenų projektavimas į

aukštesnės dimensijos erdvę, pavadintą *požymių erdve* ir hiperplokštumos, atskiriančios šias dvi klases, sukūrimas šioje erdvėje. Hiperplokštuma konstruojama maksimizuojant atstumą tarp dviejų artimiausių mokymo duomenų, priklausančių dviem skirtingoms klasėms. AVM reikalauja, kad mokymo ar atpažinimo pavyzdžiai būtų atstovaujami fiksuoto ilgio vektoriais.

AVM yra binarinis klasifikatorius, kuris daro sprendimus pagal tiesinę sprendimo ribą arba hiperplokštumą, optimaliai atskiriančią dvi klases.

2.5.4. Dirbtinių neuronų tinklai

Dirbtinių neuronų tinklai (DNT) (Simpson 1990) yra vieni iš naujesnių požymių klasifikavimo įrankių. Pirmieji DNT matematiniai modeliai buvo sukurti 1943 metais Mcculloch ir Pitts. Buvo įrodyta, kad naudojant neuroną su slenkstinėmis aktyvavimo funkcijomis, galima įvykdyti bet kokią loginę operaciją (Navakauskas 2000). 1949 metais Hebb pirmą kartą suformulavo dirbtinių neuronų mokymo algoritmą (Hebb 1949), o 1957 Rosenblatt išleido pirmą veikalą apie DNT, skirtą perceptronų tinklo analizei (Rosenblatt 1957).

Šiuo metu DNT labai plačiai taikomi įvairiose srityse (Navakauskas 2000):

- Kalbos apdorojime.
- Kalbančiojo atpažinime.
- Raidžių atpažinime.
- Objektų atpažinime.
- Signalų apdorojime ir t. t.

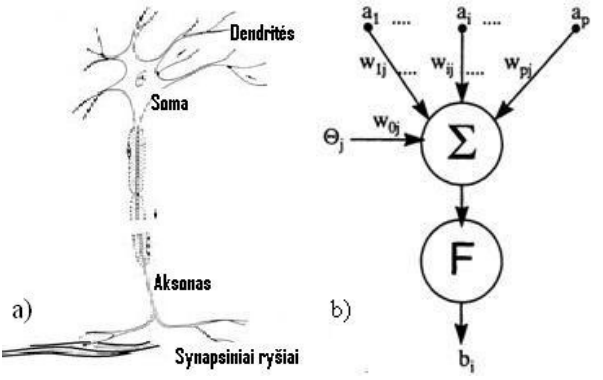
Biologiniai ir dirbtinių neuronų tinklai

Teigiama, kad žmogaus smegenyse yra 10^{11}–10^{14} neuronų, kurių kiekvienas turi apie 10^4 ryšių su kitais neuronais. Žmogaus smegenyse nėra centro, kuris valdytų visus neuronus, kiekvienas neuronas veikia nepriklausomai, naudodamas lokalią informaciją, gaunamą iš kitų neuronų. DNT kūrimo tikslas nėra kopijuoti žmogaus ar gyvūno smegenis, bet perimti neuronų sąveikos mechanizmus efektyvesniam informacijos apdorojimui (Navakauskas 2000).

Biologinio neurono supaprastintas modelis pateiktas 2.15 paveiksle a). Įėjimo sluoksnyje yra dendritės, per kurias neuronas gauna informaciją iš kitų neuronų. Ši informacija „surenkama“ į somą ir aksonu toliau perduodama per aksonos atšakas – synapsinius ryšius į kitus neuronus.

2.15 paveiksle b) pateiktas labai supaprastintas biologinio neurono modelis – dirbtinis neuronas. Dirbtinio neurono išėjimas formuojamas kiekvieną iš įėjimo signalų padauginus iš atitinkamo koeficiento (svorio) ir susumuojant. Iš gautos sumos atimamas slenkstis ir gautas signalas paveikiamas tam tikros aktyvavimo funkcijos F:

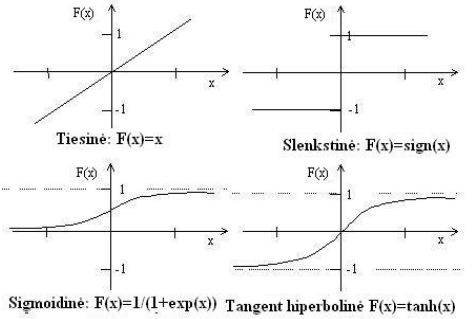
$$b_j = \mathrm{F}\left(\sum_{i=1}^{p} a_i w_{ij} - w_{0j}\Theta_j \right). \qquad (3.86)$$



2.15 pav. a) biologinio neurono modelis; b) dirbtinis neuronas (Navakauskas 2000)

Fig. 2.15. a) model of biologic neuron; b) artificial neuron (Navakauskas 2000)

Kai kurios neuronų aktyvavimo funkcijos F pateiktos 2.16 paveiksle.



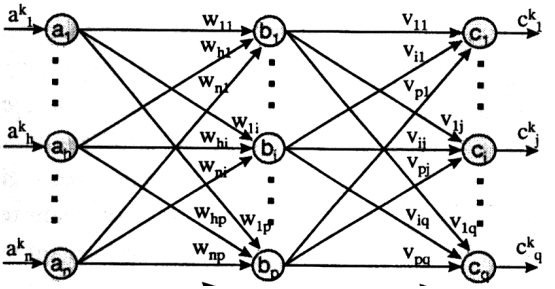
2.16 pav. Kai kurios neuronų aktyvavimo funkcijos

Fig. 2.16. Some activation functions of the neurons

Jungiant neuronų grupes (sluoksnius) tarpusavyje, gaunami neuronų tinklai. Jungimo schemų gali būti įvairių, jos skirstomos pagal ryšius tokiu būdu:

- Intra-sluoksniniai ryšiai (jungia to paties sluoksnio neuronus).
- Inter-sluoksniniai ryšiai (jungia skirtingų sluoksnių neuronus).

Juos dar galima skirstyti į tiesioginius, grįžtamuosius ir rekurentinius.



2.17 pav. Dirbtinio neuronų tinklo pavyzdys (Navakauskas 2000)

Fig. 2.17. Artificial neural network (Navakauskas 2000)

2.17 paveiksle pateiktas dirbtinių neuronų tinklo pavyzdys. Šis tinklas turi tris sluoksnius su tiesioginiais inter-sluoksniniais ryšiais. Pirmas sluoksnis, turintis tiesioginius ryšius su įėjimais, vadinamas įėjimo sluoksniu. Paskutinis sluoksnis, turintis tiesioginį ryšį su išėjimais, vadinamas išėjimo sluoksniu. Visi kiti sluoksniai vadinami paslėptaisiais (šiame pavyzdyje yra vienas paslėptasis sluoksnis). Pateiktame pavyzdyje tinklas įėjimo sluoksnyje turi *n* neuronų, paslėptame – *p*, o išėjimo sluoksnyje – *q* neuronų.

DNT mokymas

Vienas iš svarbiausių etapų, naudojant DNT, yra jų mokymas. Mokymo algoritmai priklauso nuo pasirinkto tinklo struktūros. Mokymo metu keičiami ryšio (svorių) koeficientai w_{ij}, bei slenksčiai Θ_i. Toks mokymas gali būti su mokytoju ir be mokytojo. Jei DNT mokymo metu žinomi mums reikiami DNT atsako duomenys, toks mokymas vadinamas su mokytoju. Kai vykdomas mokymas su mokytoju, į DNT įėjimą pateikiami mokymui skirti duomenys (požymių vektoriai) ir pagal gautus DNT išėjimus, naudojant tam tikrus algoritmus, keičiami ryšio (svorių) koeficientai, kad galėtume gauti DNT išėjimą tokį, kokį mes norime. DNT gali būti heteroasociatyvūs ir autoasociatyvūs. Heteroasociatyvių DNT atveju, norimas tinklo išėjimas yra kitoks nei įėjimas. Autoasociatyvių DNT atveju, mokymo metu tinklas savo išėjime bando atkartoti vektorių, kuris paduotas į tinklo įėjimą. Mokymo be mokytojo atveju, DNT atsakas į pateiktą įėjimą nėra žinomas, todėl DNT tenka pačiam jo ieškoti. Šia mokymo rūšimi pasižymi, pvz. save organizuojantys DNT (Kohoneno DNT). Tokio mokymo metu gali būti naudojamas stochastinis mokymas, kada atsitiktinai keičiami DNT ryšių svoriai ir stebimas dėl šio ryšių svorio pakeitimo įvykęs DNT energijos pokytis: jei ji pasikeičia, ryšių svoris gali būti pakeistas,

priešingu atveju, ryšių svoris paliekamas koks buvo. Kitas mokymo be mokytojo realizavimo būdas gali būti, kai į pateikus pavyzdį į DNT įėjimą, išėjimo sluoksnio neuronai varžosi per rekurentinius ryšius siųsdami sau teigiamus signalus, o kaimyniniams šio sluoksnio neuronams – neigiamus. Nusistovėjus pusiausvyrai, išėjime lieka vienas aktyvus neuronas.

Kadangi DNT veikia kaip klasifikatoriai, juos galima panaudoti kalbančiojo atpažinimui. Šiam tikslui gali būti panaudoti tiek autoasociatyvūs, tiek ir heteroasociatyvūs DNT (Badran, Selim 2000; Farrell *et al.* 1994; Fredrickson, Tarassenko 1995; Hattori 1992; Mary *et al.* 2004).

2.6. Kalbančiojo atpažinimo sistemų pavyzdžiai

Kalbančiojo atpažinimo sistemų pasaulyje sukurta daug ir įvairių. Šios sistemos kuriamos įvairiai derinant požymių vektorių sistemas ir klasifikavimo metodus. Yra sprendžiamos robastinio (patikimo) asmens atpažinimo problemos ir t. t. Trumpai paminėsime keletą iš sukurtų naujesnių kalbančiojo atpažinimo sistemų.

Nepriklausoma nuo ištarto teksto kalbančiojo atpažinimo sistema, gauta derinant fonetinius, idiolektinius ir akustinius požymius, aptariama (Andrews *et al.* 2001). Šios sistemos lygių klaidų lygis pasiekiamas mažesnis nei 1 %. Fonetinė sistema yra nauja nuo kalbos nepriklausanti kalbančiojo atpažinimo sistema, naudojanti labiau fonetinius (pvz. tartis) požymius nei spektrinius. Sistema naudoja šešių kalbų fonetinę informaciją atpažinti kalbančiajam. Idiolektinė sistema modeliuoja kalbančiojo išskirtinius bruožus, skaičiuodama žodžių dažnumą, panaudojant automatinę kalbos atpažinimo sistemą. Akustinė sistema naudoja kalbos signalo spektrinius skirtumus. Čia klasifikavimui panaudojami Gauso mišinių modeliai.

Kalbančiojo verifikavimo sistema, kurioje sprendžiamos robastinio kalbančiojo atpažinimo problemos, aptariama (Wong, Russell 2001). Realiose sistemose adityvinis ar multiplikatyvinis triukšmas sukuria įvairius neatitikimus mokymo ir atpažinimo procesų metu. Šiai problemai spręsti naudojama lygiagreti modelių kombinacija PMC (angl. – *Parallel Model Combination*). Aptariama PMC taikymas priklausančiame nuo pasakyto teksto kalbančiojo atpažinime, panaudojant PMM. Kalbos ir triukšmo duomenys gauti iš YOHO ir NOISEX-92 duomenų bazių.

Požymių sistemos, susidedančios iš žadinimo signalo pagrindinio dažnio, formančių centrinių dažnių ir formančių pločių panaudojimas kalbančiojo verifikavimo sistemoje aptartas (Hansen *et al.* 2001). Požymiai randami panaudojant entropinę signalų apdorojimo sistemą. Klasifikavimui panaudota GMM bei rezultatų palyginimui – vektorinis kvantavimas. Taip pat panaudoti

įvairūs normalizacijos metodai. Kartu, rezultatų palyginimui, atlikti eksperimentai panaudojant standartinius MSKK. Šiuo atveju, panaudojant MSKK, buvo gauti tikslesni atpažinimo rezultatai.

Aukštesnės eilės spektro fazės požymių panaudojimas kalbančiojo verifikavimo sistemoje aptartas (Ning, Chandran 2004). Paprastai fazinė spektro informacija atmetama ir laikoma nesvarbia kalbančiojo atpažinime. Didžioji dalis požymių randama iš amplitudinio spektro. Šioje sistemoje panaudota tiek amplitudinė, tiek ir fazinė spektro informacija. Klasifikavimui panaudota GMM. Panaudojant aukštesnės eilės spektro fazinius požymius gautas atpažinimo tikslumas artimas MSKK. Fazinių požymių efektyvumas buvo parodytas atlikus eksperimentus iš kalbos signalų pašalinus amplitudinę spektrinę informaciją. Taip pat buvo parodyta, kad šie faziniai požymiai yra atsparesni adityviniam baltam Gausiniam triukšmui, kuomet įrašymo ir atpažinimo sąlygos kitokios nei MSKK**.**

Kalbančiojo atpažinimo sistema, naudojanti balsaskylės modelius ir uždaros fazės analizę, aptarta (Slyh *et al.* 2004). Balsaskylės modelis pirmą kartą buvo pasiūlytas Fujisaki ir Ljungqvist ir buvo sujungtas su uždaros fazės analize tam, kad būtų galima gauti požymius, naudojamus kalbančiojo atpažinimui. Taipogi šie požymiai buvo apjungti kartu su sistemos požymiais, naudojančiais formančių centrinius dažnius, formančių pločius bei F0 (FMBWF0). Paaiškėjo, kad apjungus šiuos požymius, gauti geresni rezultatai, nei panaudojant tik FMBWF0, ir atpažinimo tikslumu pralenkė standartinius MSKK. Klasifikavimui panaudota GMM.

Kalbančiojo atpažinimo sistemos, gautos panaudojant priklausomus nuo fonemų Gauso mišinių modelius, aptartos (Hansen *et al.* 2004). Aptarti trys būdai tokių GMM sudarymui. Bet kokios sistemos, aprašančios GMM atskirą fonemą, veikimas prastesnis nei standartinės GMM sistemos, aprašančios bendrai visas fonemas. Tačiau apjungus atskirų fonemų sistemas gauti rezultatai 2,6 % pralenkė standartinį metodą.

Kalbančiojo atpažinimo sistema, panaudojanti priklausomus nuo kalbėtojo požymius – kepstro koeficientus. Šie požymiai randami iš vidurkinio normuoto TPM spektro, kuris apskaičiuojamas vidurkinant autokoreliacinės funkcijos koeficientus ir po to juos normuojant. Iš šių koeficientų pritaikius Levinsono – Durbino algoritmą randamas vidurkinis normuotas TPM spektras. Pagal šio spektro gaubtinę suformuojamas specialus filtrų rinkinys bei apskaičiuojamas spektras, panašiai kaip ir skaičiuojant melų skalės spektro koeficientus (MSSK). Toliau pritaikius diskrečiąją kosinusų transformaciją gaunami priklausantys nuo kalbėtojo kepstro koeficientai. Kalbėtojo modeliavimui panaudoti tolydinio tankio ergodiniai paslėptieji Markovo modeliai (Orsag 2004).

Kalbančiojo atpažinimo sistema, naudojanti požymių klasifikavimui atraminių vektorių mašinas (AVM), aptarta (Campbell *et al.* 2007). Kaip jau

buvo minėta, šiuo metu yra tendencija naudoti aukšto lygio požymius, tokius kaip tartis, žodžių naudojimas, prozodija ir t. t. Aukšto lygio sistemos turi labai daug duomenų charakterizuoti kalbantįjį. Šiuo metu daug pastangų dedama aukšto lygio požymių paieškai, tačiau mažiau pastangų dedama šių požymių modeliavimui. Pasiūlyta modeliuoti kalbantįjį panaudojant AVM, panaudojant naujus branduolius, tiesinančius logaritminę tikėtinumo santykio įverčio sistemą. Parodytas šio metodo efektyvumas daugeliui aukšto lygio požymių, taip pat prieš standartinį logaritminį tikėtinumo santykio modeliavimą. Taip pat įrodyta, kad ši sistema gali būti efektyviai sujungta su standartine kalbančiojo atpažinimo sistema.

2.7. Antrojo skyriaus apibendrinimas ir disertacijos uždavinių formulavimas

- Kalbos signalų generavimas yra labai sudėtingas procesas, kurį tiksliai matematiškai aprašyti neįmanoma. Dėl to kuriami supaprastinti kalbos generavimo modeliai ir pagal šių modelių parametrus nustatomi tam tikri kalbėtojų požymiai, kurie naudojami atpažinime.
- Dažniausiai tie patys požymiai naudojami tiek kalbos, tiek ir asmens atpažinimo sistemose, nors tai yra du skirtingi uždaviniai. Iš spektrinių požymių šiuo metu bene populiariausi yra melų skalės kepstro koeficientai (MSKK).
- Kalbėtojų modeliavimui ir požymių vektorių palyginimui yra sukurta daug metodų, kiekvienas iš jų turi tam tikrų privalumų ir trūkumų. Panaudojant paprastesnius ir lengviau realizuojamus metodus, dažniausiai gaunami prastesni atpažinimo rezultatai.
- Kuriant kalbančiojo atpažinimo sistemas reikia spręsti tokius klausimus, kaip tinkamos požymių sistemos ir jų klasifikacijos metodo parinkimo, kalbos signalų atskyrimo nuo triukšmo ir t. t.
- Disertacijos uždaviniai – atlikus kalbančiojo atpažinimo sistemų analizę, paminėjus tam tikras problemas, su kuriomis susiduriama kalbančiojo atpažinime, pasiūlyti tam tikrus sprendimus, jų pagrindu sukurti atpažinimo sistemą, bei atlikti jos eksperimentinį tyrimą:
    1. Kadangi kalbančiojo atpažinimo sistemose naudojamas „kalbos detektorius“, išrenkantis kalbos signalus iš triukšmo, o tam tikslui nėra sukurtų matematiškai pagrįstų metodų, vienas iš darbo uždavinių –

pasiūlyti vokalizuotų garsų atskyrimo nuo triukšmo metodą, kuris būtų greitas ir paprastai realizuojamas, be to, pilnai automatizuotas.

2. Kadangi tos pačios požymių sistemos naudojamos tiek asmens, tiek ir kalbos atpažinimo sistemose ir iki šiol nėra rasta požymių, kurie vienareikšmiškai apibūdintų kalbantįjį asmenį, ir nepriklausytų nuo kalbos, vienas iš darbo uždavinių – pasiūlyti naują požymių vektorių sistemą, gautą apjungus tiek žadinimo signalo, tiek balso trakto parametrus.
3. Kadangi kalbančiojo modeliavimui ir požymių vektorių palyginimui naudojant Gauso mišinių modelius viena iš problemų yra pradinis GMM parametrų vertinimas, prieš naudojant matematinės vilties maksimizavimo algoritmą, vienas iš darbo uždavinių – pasiūlyti tam tikslui nesudėtingą, greitai veikiantį ir patogų metodą.
4. Sukurti kalbančiojo atpažinimo sistemą, kurioje būtų realizuoti pasiūlyti sprendimai, bei atlikti jos eksperimentinį tyrimą. Siekiant ištirti pasiūlytos požymių sistemos efektyvumą, gautus atpažinimo rezultatus palyginti su atpažinimo rezultatais, gautais panaudojus tradicinius požymius – MSKK.

3

Atpažinimo sistemos kūrimas

Šiame skyriuje pristatysime sukurtą atpažinimo sistemą su joje realizuotais pasiūlytais sprendimais. Pristatysime pasiūlytą vokalizuotų garsų atskyrimo nuo triukšmo bei nevokalizuotų garsų (segmentavimo) algoritmą, požymių vektorių sistemą, jos ypatumus.

Kaip kalbančiojo modeliavimo ir požymių palyginimo metodą pasirinkome Gauso mišinių modelius (GMM). Pasirinkimą lėmė tikslas sukurti nepriklausomą nuo ištarto teksto kalbančiojo atpažinimo sistemą su efektyviu požymių palyginimo algoritmu. Kalbos signalo požymius pasirinkome dvejopus: standartinius melų skalės kepstro koeficientus (MSKK) ir požymius, sudarytus iš žadinimo signalo pagrindinio dažnio bei formančių ir antiformančių (jų skaičių ir kombinaciją galima pasirinkti). Vienas iš pagrindinių tikslų – yra palyginti pasiūlytą požymių sistemą su bene plačiausiai naudojama pasaulyje (MSKK), pagal atpažinimo tikslumą bei reikalingą skaičiavimo operacijų kiekį.

Skyriaus tematika paskelbti du autoriaus straipsniai (Kamarauskas 2006; Kamarauskas 2008).

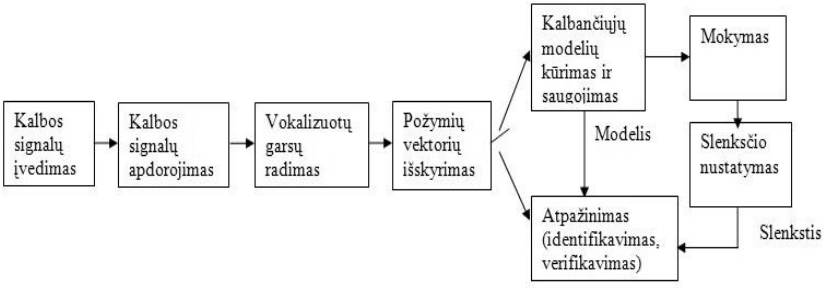
3.1. Kalbančiojo atpažinimo sistema

Darbo metu sukurta automatinė kalbančiojo atpažinimo sistema *gmm.exe*. Šioje sistemoje realizuotos visos kalbančiojo atpažinimo algoritmų grupės: atviros bei uždaros aibės kalbančiojo identifikavimas bei kalbančiojo verifikavimas. Sistema pritaikyta efektyviai dirbti su balsų bazėmis, automatiškai atidaro visas kataloge esančias garso rinkmenas, vykdo pasirinktą funkciją (mokymą, identifikavimą ar verifikavimą). Atpažinimo rezultatai pateikiami lentelėse ir grafikuose. Taip pat sistema pateikia visas jos darbo efektyvumo vertinimui reikalingas kreives: intraindividualius ir interindividualius iškraipymus, DET, KPL-KAL. Galima pasirinkti įvairius kalbos signalo apdorojimo, segmentavimo ir kitus parametrus, norimą požymių vektorių sistemą ir t. t.

Ši sistema gali būti taikoma asmenų paieškai duomenų bazėse, o taip pat kalbančiojo atpažinimo tyrimams, nes joje galima keisti daug parametrų ir tirti jų įtaką atpažinimo tikslumui.

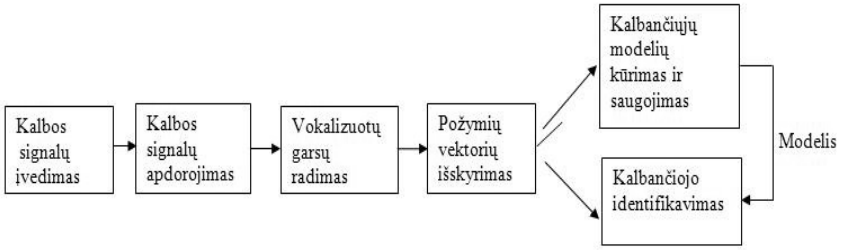
3.2. Kalbančiojo atpažinimo algoritmas

Kaip jau buvo minėta, sistemoje realizuoti visi kalbančiojo atpažinimo algoritmų tipai: identifikavimas ir verifikavimas. Visais atvejais sistemos veikimo algoritmas yra tas pats, tik atviros aibės kalbančiojo identifikavimo ir verifikavimo atveju papildomai turi būti įvykdytas mokymas ir nustatytas slenkstis. Kalbančiojo atpažinimo algoritmų blokinės schemos pavaizduotos 3.1 ir 3.2 paveiksluose.



3.1 pav. Atviros aibės kalbančiojo identifikavimo ir kalbančiojo verifikavimo algoritmo struktūrinė schema

Fig. 3.1. Algorithm of the open-set speaker identification and speaker verification



3.2 pav. Uždaros aibės kalbančiojo identifikavimo algoritmo struktūrinė schema

Fig. 3.2. Algorithm of the closed-set speaker identification

Bendra atpažinimo sistemos algoritmo schema parodyta 3.3 paveiksle.

Pirmasis atpažinimo proceso etapas yra kalbos signalo įvedimas. Toliau šis kalbos signalas apdorojamas, segmentuojamas ir iš jo apskaičiuojami požymių vektoriai, kurie gali būti panaudoti tiek modelių kūrimui, tiek ir atpažinimui.

Toliau aptarsime šių algoritmų atskirus etapus.

3.2.1. Kalbos signalų įvedimas

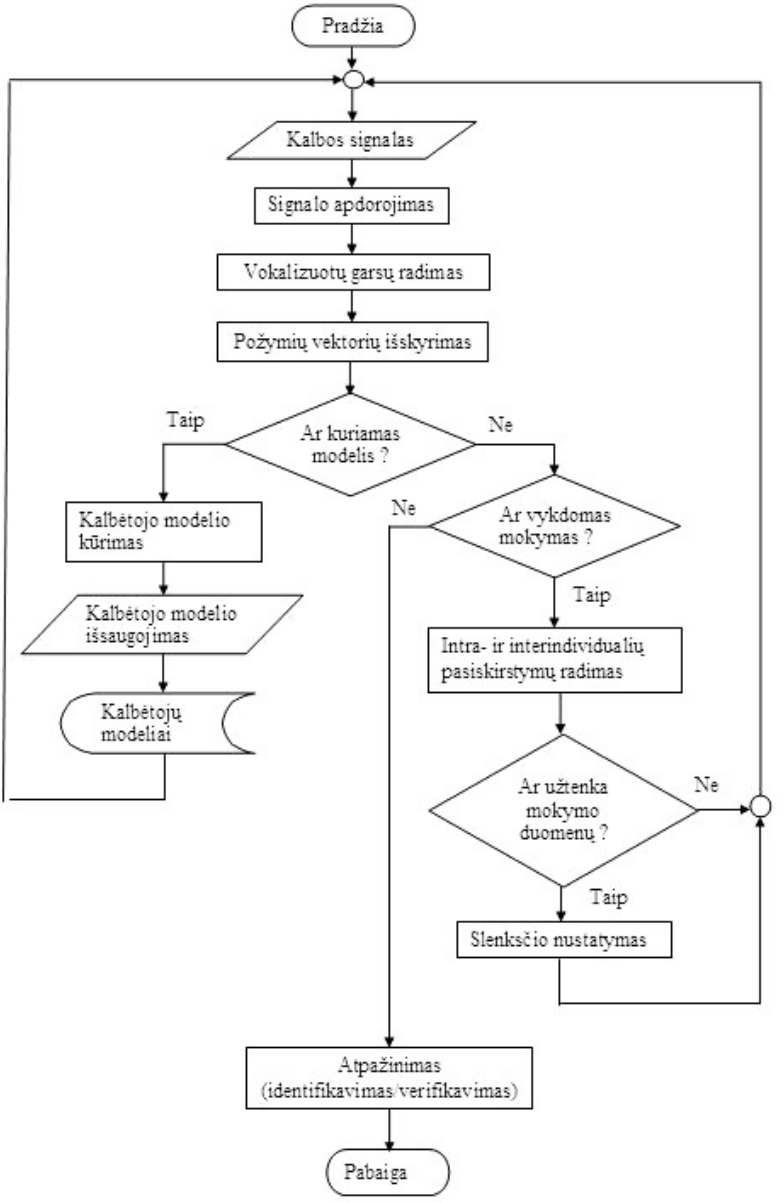
Atpažinimo sistemoje kalbos signalas nuskaitomas iš garsinės rinkmenos. Rinkmenos formatas turi būti Windows PCM (wav tipo), atskaitos koduotos 16 bitų, vienas įrašymo kanalas (mono įrašas). Geriausiai, kad garso įrašai būtų įrašyti 11 025 Hz diskretizacijos dažniu, kadangi kalbos signalo spektras išsidėstęs maždaug iki 5 kHz. Naudojant garso įrašus su žemesniu diskretizacijos dažniu įrašomas kalbos signalas, turintis siaurą dažnių juostą, aukštesnės formantės gali būti iškraipytos. Tuo tarpu esant aukštam diskretizacijos dažniui (pvz. 44 100 Hz) įrašyto kalbos signalo spektras užima nedidelę spektro dalį, aukštesniuose dažniuose išsidėstęs triukšmo spektras, dėl ko taip pat prastėja sistemos atpažinimo rezultatai.

3.2.2. Kalbos signalų apdorojimas

Kalbos signalų apdorojimas susideda iš trijų etapų:

1. Nuolatinės dedamosios atėmimas.
2. Filtravimas aukštų dažnių filtru (pradinė filtracija).
3. Signalo dalijimas į kadrus.

Toliau trumpai aptarsime šiuos etapus.



3.3 pav. Kalbančiojo atpažinimo sistemos veikimo algoritmas

Fig. 3.3. Algorithm of the speaker recognition system

Signalo **nuolatinės dedamosios pašalinimas** sumažina įrašymo įrangos įtaką tolimesnei signalo analizei. Tam tikslui randamas signalo reikšmių vidurkis ir kiekviena signalo reikšmė atimama iš šio vidurkio:

$$\widetilde{s}(n) = s(n) - \frac{1}{N}\sum_{i=1}^{N} s(n), \tag{3.1}$$

čia N – signalo bendras atskaitų skaičius, $\widetilde{s}(n)$ – apdorotas signalas, $s(n)$ – pradinis signalas.

Pradinė filtracija atliekama panaudojant pirmos eilės aukštų dažnių RIR filtrą ir yra skirta spektro išlyginimui. Ji atliekama pagal (2.1) formulę. Filtracijos koeficiento reikšmė pagal nutylėjimą pasirinkta α=0,96.

Toliau atliekamas **dalijimas į kadrus** pagal (2.3) formulę. Yra naudojami du skirtingi kadrai, vienas požymių vektorių išskyrimui, kitas žadinimo signalo pagrindinio dažnio išskyrimui. Požymių išskyrimui skirto kadro ilgis pagal nutylėjimą lygus 20 ms, o žadinimo signalo pagrindiniam dažniui – 40 ms. Kadro postūmis imamas 5 ms. Visi šie parametrai gali būti pakeisti.

3.2.3. Vokalizuotų garsų išskyrimas

Vokalizuotų garsų išskyrimas yra svarbi kalbos/kalbančiojo atpažinimo sistemų dalis. Reikėtų paminėti, kad iki šiol tam tikslui nėra sukurta matematiškai pagrįstų būdų, daugelis automatinio segmentavimo būdų remiasi euristiniais metodais. Galima paminėti vieną iš pasiūlytų segmentavimo metodų, naudojantį autoregresinių atsitiktinių sekų su šuoliais besikeičiančiais parametrais, naudojant maksimalaus tikėtinumo ir vidutinės kvadratinės klaidos kriterijus, optimalią segmentaciją (Lipeika 2000). Šiame metode tikslo funkcija yra modifikuota taip, kad optimizacijai galima būtų panaudoti dinaminio programavimo metodą. Šiai tikslo funkcijai gautos patogios taikymui Belmano funkcijų išraiškos. Tačiau šis metodas yra pakankamai sudėtingas, be to nėra pilnai automatizuotas. Dar galima būtų paminėti ir kitų pasiūlytų segmentavimo būdų: statistinį ištisinės kalbos segmentavimo metodą (Obrecht 1988), kur signalas modeliuojamas statistiniu modeliu ir bandoma rasti modelio parametrų pasikeitimus. Taip pat pasiūlytą kitą statistinį segmentavimo metodą, skirtą sujungtų žodžių atpažinimo sistemai, kur vertinimo procedūroje, nustatančioje žodžio ribas, naudojami kvadratiniai polinomai (Zelinski, Class 1983). Yra sukurta ir daugiau įvairių metodų: tokių kaip kalbėtojo modeliavimas triukšmingoje aplinkoje (Toledo-Ronen 2001), „kalbos detektorius“, naudojantis vokalizavimo lygį bei energiją, kur skaičiuojant vokalizavimo lygį naudojama

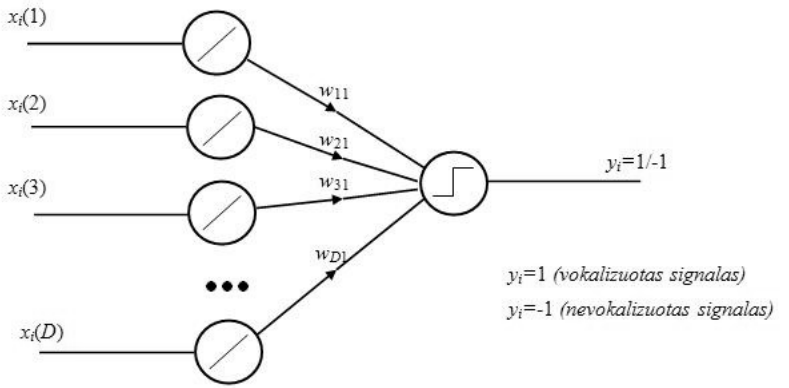
netiesinė centro iškirpimo funkcija (Zilca *et al.* 2004), „kalbos detektorius“, naudojantis dviejų būsenų (atitinkančių kalbą ir triukšmą) paslėptuosius Markovo modelius (Ore *et al.* 2006) ir daug kitų.

Panaudojant nesudėtingus klasikinius metodus, tokius kaip energijos slenksčio nustatymas, gaunami prasti segmentavimo rezultatai, ypač esant aukštesniam triukšmo lygiui. Be to, kai kuriems tokiems metodams reikia nurodyti kalbos signalo ir triukšmo pavyzdžius, o tai yra nepriimtina realiose atpažinimo sistemose.

Darbo metu mes pasiūlėme du vokalizuotų garsų išrinkimo iš įrašytų kalbos signalų metodus. Vienas – iš „kalbos detektorių“ realizuotas, panaudojant dirbtinių neuronų tinklus (DNT) (Kamarauskas 2006).

3.2.3.1. Vokalizuotų garsų išrinkimas taikant dirbtinių neuronų tinklus

Tam tikslui panaudojome dviejų tipų DNT: *perceptronų* ir *atgalinio sklidimo* (angl. – *back propagation*). Panaudota perceptronų tinklo struktūra pateikta 3.4 paveiksle.



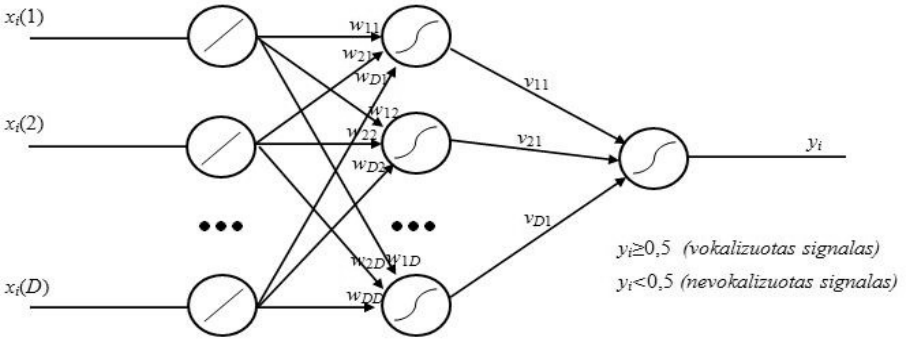
3.4 pav. Perceptronų tinklo struktūra, panaudota vokalizuotų garsų išrinkimui

Fig. 3.4. Structure of the perceptrons ANN used for the segmentation of the voiced sounds

Kaip matome, šis perceptronų DNT susideda iš dviejų sluoksnių. Įėjimo sluoksnyje panaudoti dirbtiniai neuronai su tiesinėmis aktyvavimo funkcijomis, jų skaičius lygus naudojamų požymių vektorių komponenčių skaičiui. Kadangi reikia atskirti dviejų tipų signalus: vokalizuotus ir nevokalizuotus, išėjimo

sluoksnyje panaudojome vieną neuroną su slenkstine aktyvavimo funkcija. Šis DNT buvo mokomas panaudojant standartinę perceptrono klaidos korekcijos procedūrą (Navakauskas 2000) taip, kad į DNT įėjimą padavus signalo kadro požymių vektorių, esant vokalizuotam signalo kadrui, tinklo išėjimo signalas būtų lygus 1 (t. y. y=1), esant nevokalizuotam, tinklo išėjime būtų −1(t. y. y=−1).

Panaudota atgalinio sklidimo DNT struktūra parodyta 3.5 paveiksle.



3.5. pav.Atgalinio sklidimo DNT struktūra, panaudota vokalizuotų garsų išrinkimui

Fig. 3.5. Structure of the back-propagation ANN used for the segmentation of the voiced sounds

Kaip matome, šis atgalinio sklidimo DNT susideda iš trijų sluoksnių. Įėjimo sluoksnyje panaudoti dirbtiniai neuronai su tiesinėmis aktyvavimo funkcijomis, jų skaičius lygus naudojamų požymių vektorių komponenčių skaičiui. Paslėptojo (vidurinio) sluoksnio neuronų skaičius lygus įėjimo sluoksnio neuronų skaičiui. Čia panaudoti neuronai su sigmoidine aktyvavimo funkcija. Išėjimo sluoksnyje taip pat panaudojome vieną neuroną su sigmoidine aktyvavimo funkcija. Šis DNT, panaudojant standartinį atgalinio sklidimo mokymo algoritmą (Navakauskas 2000), buvo mokomas taip, kad į DNT įėjimą padavus požymių vektorių, esant vokalizuotam signalo kadrui, tinklo išėjimo signalas būtų lygus 0,5 (t. y. y=0,5), esant nevokalizuotam, tinklo išėjime būtų 0 (t. y. y=0). Segmentavimo metu, klasifikuojant požymių vektorius, jei tinklo išėjimas $y \geq 0,5$, (t. y. $0,5 \leq y < 1$), kadras priskiriamas vokalizuotiems, jei $y < 0,5$ (t. y. $0 < y < 0,5$), kadras laikomas nevokalizuotu.

Kalbos signalų kadrų klasifikavimui panaudotos dvi požymių sistemos: TPM parametrai bei Furjė spektro koeficientai.

Tinkamai atlikus šių DNT mokymą pasiekiami geri segmentavimo rezultatai, tačiau šis metodas turi ir tam tikrų trūkumų. Visų pirma, tai nėra pilnai

automatizuotas metodas, kadangi reikia atlikti šių tinklų mokymą, pateikiant jiems vokalizuotų garsų bei triukšmo pavyzdžius. Jeigu visuose garso įrašuose foninis triukšmas yra panašus ir pakankamai žemo lygio, galima vieną kartą atlikus šių tinklų mokymą, segmentaciją atlikti visiems garso įrašams. Tačiau, jei skiriasi garso įrašymo sąlygos, tiems įrašams gali tekti atskirai mokyti DNT. Tai nėra patogu vartotojui. Pilnai automatizuoti šį metodą yra pakankamai sudėtinga.

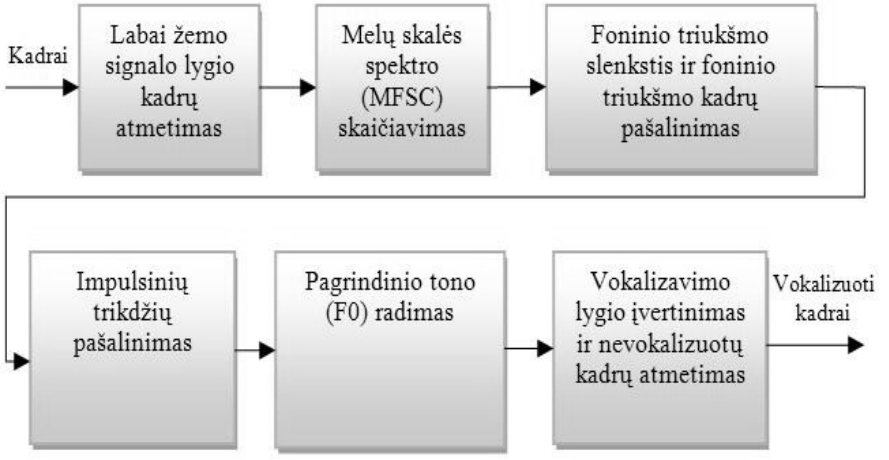
Pagrindiniai mūsų keliami reikalavimai „kalbos detektoriui“:

- Algoritmo paprastumas.
- Greitas veikimas.
- Atrenkami tik vokalizuoti garsai.
- Automatinis veikimas, nereikalaujantis iš vartotojo jokių papildomų veiksmų.

3.2.3.2. Vokalizuotų garsų išrinkimas automatiškai nustatant triukšmo parametrus

Įvertinę įvairių naudojamų vokalizuotų garsų išrinkimo metodų privalumus ir trūkumus bei suformulavę pagrindinius reikalavimus „kalbos detektoriui“, mes pasiūlėme kitą, pakankamai nesudėtingą, automatinį vokalizuotų garsų išrinkimo iš kalbos signalų metodą.

Šio algoritmo struktūrinė schema pateikta 3.6 paveiksle.



3.6 pav. Pasiūlyta automatinio vokalizuotų garsų išrinkimo algoritmo schema

Fig. 3.6. Scheme of proposed voice activity detection algorithm

Labai žemo signalo lygio kadrų atmetimas

Tokių kadrų, kuriuose nėra signalo (signalo reikšmės nulinės ar labai žemos) analizė neturi jokios prasmės, kadangi nėra jokio kalbos signalo, o įrašytos nulinės reikšmės arba analoginio – skaitmeninio keitiklio (ASK) žemiausių skilčių triukšmai. Tam tikslui kiekviename kadre ieškoma maksimali signalo amplitudės modulio reikšmė ir ji lyginama su slenksčio reikšme, kuri pagal nutylėjimą lygi 130. Jei ši reikšmė mažesnė už slenkstį, šis kadras tolesniuose skaičiavimuose nebenaudojamas:

$$S(j) = \max_{1 \le n \le N} (|\, s(j,n)\,|)\,, \tag{3.2}$$

čia j=1, 2,…, K, n=1, 2,…, N, kur K – kadrų skaičius, N – kadro ilgis, funkcija max(x) gražina masyvo x didžiausią reikšmę.

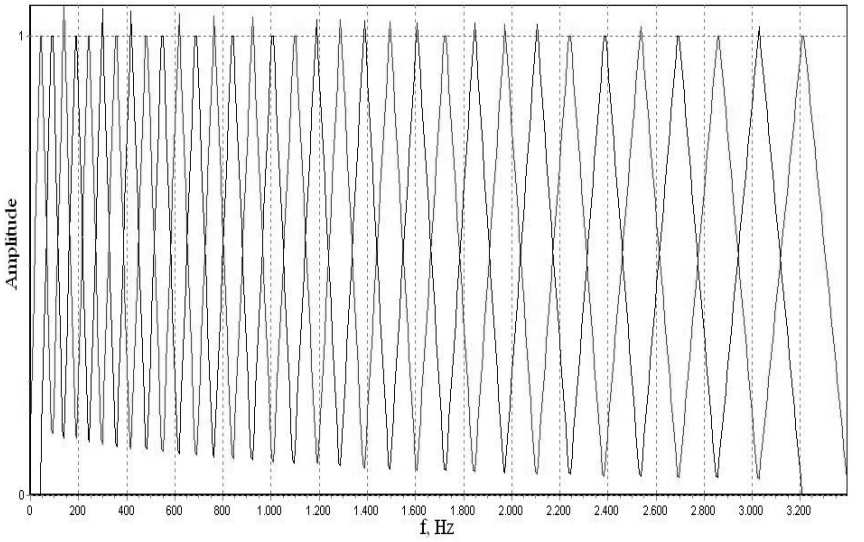
Sekančiame žingsnyje likusiems kalbos signalo kadrams naudojama **Hemingo lango funkcija**. Signalo kadrų dauginimas iš Hemingo lango funkcijos atliekamas pagal (2.4) ir (2.6) formules.

Melų skalės spektro skaičiavimas

Sekantis žingsnis yra melų skalės spektro skaičiavimas. Tam tikslui pradžioje apskaičiuojamas signalo kadro Furjė spektras (DFT), taikant GFT algoritmą. Spektro taškų skaičius lygus 512. Toliau pagal (2.37)–(2.41) apskaičiuojami trikampiai filtrai, išsidėstę pagal melų skalę ir randamas melų skalės spektras (MSSK). Tai atliekama kiekvienam naudojamam kadrui. Parinktas trikampių filtrų skaičius I=33.

$$E(m,i) = \sum_{k=1}^{512} \left| X_F(m,k) \right| H(i,k), \tag{3.3}$$

čia I=33, trikampių filtrų skaičius, M − naudojamų kadrų skaičius, $H(i,k)$ – trikampių filtrų funkcija, $X_F(m,k)$ − m-tojo kadro Furjė spektras. $1 \le i \le I$, $1 \le m \le M-1$. Trikampių filtrų rinkinys, panaudotas MSSK skaičiavimui, pavaizduotas 3.7 paveiksle.



3.7 pav. Trikampių filtrų išsidėstymas pagal melų skalę, kai dažnių juosta 0–3 400 Hz

Fig. 3.7. Triangular filters allocated according to mel-frequency scale, when bandwidth 0–3 400 Hz

Foninio triukšmo slenkstis ir foninio triukšmo kadrų pašalinimas

Sekantis žingsnis yra foninio triukšmo radimas. Daugelyje kalbos signalų segmentavimo sistemų siekiant rasti triukšmo parametrus, daroma prielaida, kad garso įrašo pradžioje yra foninis triukšmas, ir imamas tam tikras skaičius kadrų nuo garso įrašo pradžios bei randami reikiami triukšmo parametrai (energija ir kiti). Jei kalbos signalas prasidės nuo pat garso įrašo pradžios, tokia segmentavimo sistema darys dideles klaidas. Siekiant rasti foninio triukšmo parametrus, mes tikriname visus likusius garso įrašo signalo kadrus. Tam tikslui randama 10 kadrų, turinčių mažiausią melų skalės spektro energijos vidurkį, kurį galime išreikšti:

$$E_{av}(m) = \frac{1}{I}\sum_{i=1}^{I} E(m,i), \tag{3.4}$$

čia I=33 – melų skalės spektro (trikampių filtrų) dedamųjų skaičius.

Taigi, randama 10 minimalių $E_{av}(m)$ reikšmių. Pradžioje iš 10 kadrų, turinčių mažiausias energijos spektro vidurkių reikšmes, apskaičiuojamas foninio triukšmo kiekvienos melų skalės spektro komponentės vidurkis:

$$E_n(i) = \frac{\sum_{m=1}^{10} E(m,i)}{10}. \tag{3.5}$$

Toliau atliekamas slenksčio skaičiavimas:

$$Thr = 2 \cdot \frac{1}{I} \sum_{i=1}^{I} E_n(i). \tag{3.6}$$

Po to skaičiuojamas kiekvieno kadro melų skalės spektro vidurkis:

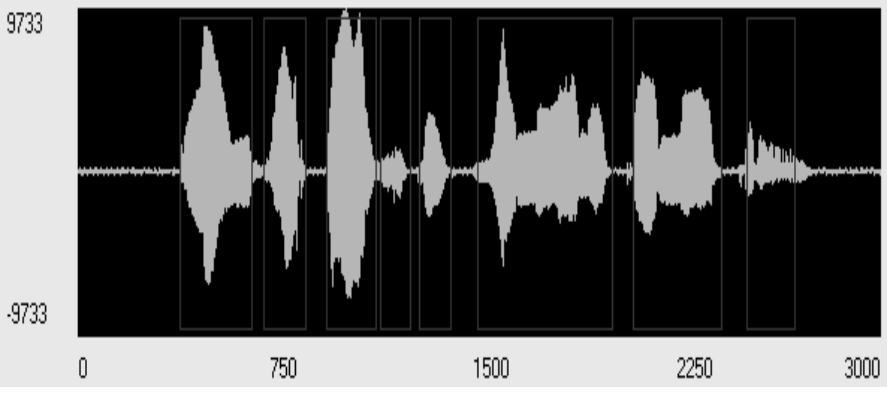
$$E_{mean}(m) = \frac{1}{I} \sum_{i=1}^{I} E(m,i), \tag{3.7}$$

bei lyginamas su slenksčiu *Thr*. Kadrai, kurių melų skalės spektro vidurkis mažesnis už slenkstį *Thr*, atmetami.

Impulsinių trikdžių pašalinimas

Vėliau yra surandami ir atmetami garsai, kurių trukmė mažesnė už 15 ms, kadangi žmogus fiziologiškai tokių trumpų garsų (balsių) ištarti negali. Tikslas – pašalinti įvairius pavienius impulsinius trukdžius ir t. t.

Aukščiau aprašyto segmentavimo algoritmo veikimas iliustruotas 3.8 paveiksle. Pateiktas susegmentuoto kalbos signalo pavyzdys. Stačiakampiais apvesti signalo segmentai, kurie bus toliau naudojami žadinimo signalo pagrindinio dažnio skaičiavimui.



3.8 pav. Susegmentuoto kalbos signalo pavyzdys

Fig. 3.8. Segmented speech signal

Pagrindinio tono (F0) radimas

Tikslus ir patikimas pagrindinio tono (arba jam atvirkštinio dydžio − žadinimo signalo pagrindinio dažnio – F0) matavimas ne visuomet yra lengvas uždavinys dėl keleto priežasčių (Rabiner *et al.* 1976). Visų pirma, balsaskylės žadinimo signalas nėra graži impulsų seka. Antra, atsiranda tam tikra sąveika tarp balsaskylės ir balso trakto. Kai kuriais atvejais formantės gali gerokai pakeisti žadinimo signalo formą taip, kad tikrą pagrindinį toną gali būti sudėtinga rasti.

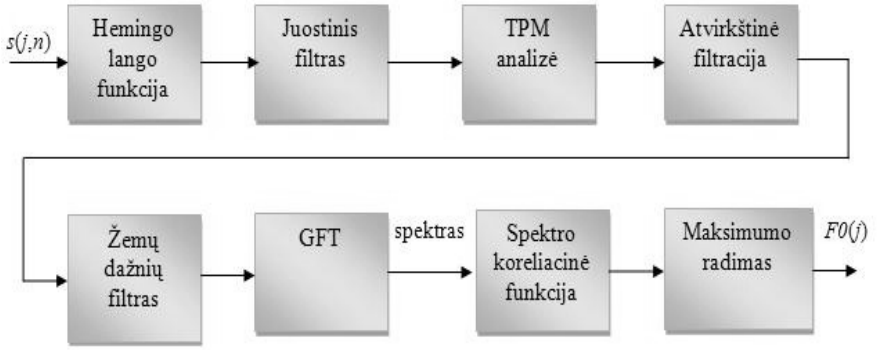
Visus pagrindinio tono radimo algoritmus galima būtų suskirstyti į tris kategorijas (Rabiner *et al.* 1976):

- Metodai, naudojantys laiko srities signalų savybes (laikiniai metodai).
- Metodai, naudojantys dažnių srities signalų savybes (dažniniai metodai).
- Metodai, naudojantys tiek laiko, tiek dažnių srities signalų savybes (kombinuoti metodai).

Įvairūs pagrindinio tono detektoriai aptarti ir palyginti (Rabiner *et al.* 1976). Dažnai naudojami F0 detektoriai, skaičiuojantys žadinimo signalo koreliacinę funkciją tiek laiko, tiek dažnių srityse. F0 radimui taip pat naudojamas ir kepstras.

Žadinimo signalo pagrindinio dažnio radimui panaudojome šiek tiek modifikuotą dažninį metodą. Dažninio metodo esmė yra žadinimo signalo pagrindinio dažnio radimas iš žadinimo signalo spektro koreliacinės funkcijos. Šio algoritmo struktūrinė schema pavaizduota 3.9 paveiksle.

Kadangi žadinimo signalas keičiasi lėčiau nei balso traktas, žadinimo signalo pagrindinio dažnio radimui naudojamas ilgesnės trukmės signalo kadras nei požymių vektorių išskyrimui. Vyriškiems balsams kadro ilgis apytiksliai lygus 40 ms, moteriškiems gali būti panaudotas ir trumpesnis (apie 25–30 ms).



3.9 pav. Žadinimo signalo pagrindinio dažnio skaičiavimo algoritmo struktūrinė schema

Fig. 3.9. Scheme of the pitch calculation algorithm

Pradžioje signalo kadras dauginamas iš *Hemingo langa* funkcijos, pagal (2.4) ir (2.6) formules.

Sekantis žingsnis yra *filtracija*. Yra naudojamas 32 eilės juostinis ribotos impulsinės reakcijos (RIR) filtras. Šio filtro pralaidumo juosta 60–3 300 Hz. Filtracija laiko srityje skaičiuojant kompoziciją tarp filtro impulsinės reakcijos ir signalo:

$$s_f(n) = s(n) * h_l(n) = \sum_{k=0}^{32} h_l(k) s(n-k), \tag{3.8}$$

čia $s_f(n)$ – filtruotas signalas, $s(n)$ – pradinis signalas, $h_l(n)$ – filtro impulsinė reakcija. Šis dažnių ruožas parinktas dėl to, kad žadinimo signalo pagrindinio dažnio apatinė riba siekia maždaug 70 Hz. Viršutinė riba parinkta tokia dėl to, kad aukštesniuose dažniuose kalbos signalo spektro intensyvumas labai mažas ir didesnė triukšmų įtaka, be to esant telefoniniams garso įrašams, jų kanalo pralaidumo juosta yra iki 3 400 Hz.

Sekantis žingsnis yra *TPM analizė*. Skaičiuojant TPM koeficientus naudojamas autokoreliacinis metodas. Pradžioje, apskaičiuojami 8 (k=0,…, 8) eilės normuoti autokoreliacinės funkcijos koeficientai:

$$r(k) = \frac{1}{N-k} \sum_{n=0}^{N-|k|-1} s(n) s(n-k), \tag{3.9}$$

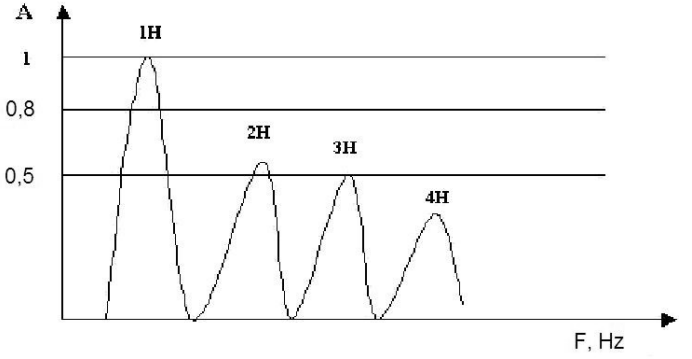
po to naudojant iteratyvinį Levinsono-Durbino algoritmą, pagal (2.21)–(2.24) formules, apskaičiuojami 8 eilės TPM modelio parametrai.

Turint TPM parametrus, galime rasti žadinimo signalą. Tam tikslui naudojama atvirkštinė filtracija. Žadinimo signalas randamas pagal (2.18) formulę. Šis žadinimo signalas sklinda iš balso stygų, jų virpėjimo periodas vadinamas *pagrindiniu tonu*. Jam atvirkštinis dydis – žadinimo signalo pagrindinis dažnis F0.

Turint žadinimo signalą, toliau atliekama jo *filtracija* 32 eilės žemų dažnių RIR filtru. Šio filtro nukirtimo dažnis ties 2 000 Hz. Imamas platesnis dažnių ruožas tam, kad gautume daugiau žadinimo signalo pagrindinio dažnio kartotinių harmonikų reikšmių.

Toliau filtruotam signalui taikoma greitoji Furjė transformacija (GFT) ir randamas jo spektras. Pagal nutylėjimą imama 1 024 spektro taškų. Tuomet, jei garso įrašo diskretizacijos dažnis F_d=11 025 Hz, gauname, kad jo spektras išsidėstęs $F_d/2$=5 512 Hz ruože. Tuomet spektro skiriamoji geba bus lygi 5 512/1 024=5,38 Hz ir žadinimo signalo pagrindinio dažnio radimo paklaida lygi 5,38/2=2,69 Hz.

Šiame spektre, kurio pavyzdys pateiktas 3.10 paveiksle, matomos žadinimo signalo pagrindinio dažnio bei kartotinės harmonikos.

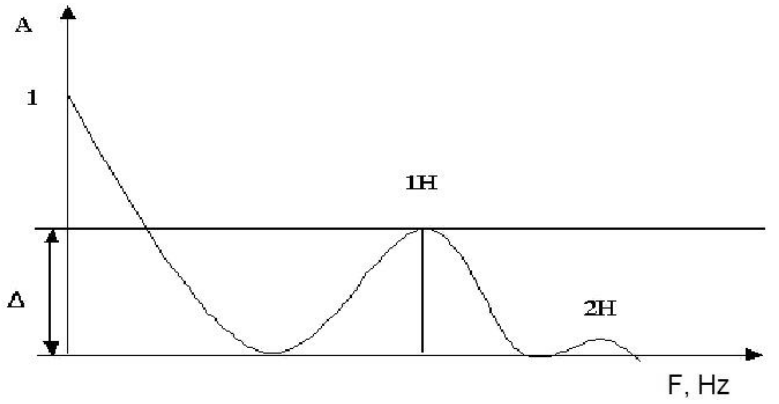


3.10 pav. Liekanos (žadinimo) signalo spektro pavyzdys

Fig. 3.10. Example of spectrum of the source signal

Siekiant rasti žadinimo signalo pagrindinį dažnį, skaičiuojama paklaidos (žadinimo) signalo spektro koreliacinė funkcija pagal (3.9) formulę. Spektro koreliacinės funkcijos pavyzdys pateiktas 3.11 paveiksle.

Toliau iš šios spektro koreliacinės funkcijos galima rasti žadinimo signalo pagrindinį dažnį (pagrindinį toną). Tuo tikslu randamas pirmasis nuo pradžių šios funkcijos maksimumas (3.11 paveiksle 1H). Šio maksimumo dažnis atitiks žadinimo signalo pagrindinį dažnį. Dažniausiai randamas maksimumas ir lyginamas su slenksčiu Δ (3.11 paveikslas), jei jis viršija šį slenkstį, (pvz. Δ=0,25), laikoma, kad šis maksimumas atitinka žadinimo signalo pagrindinį dažnį.



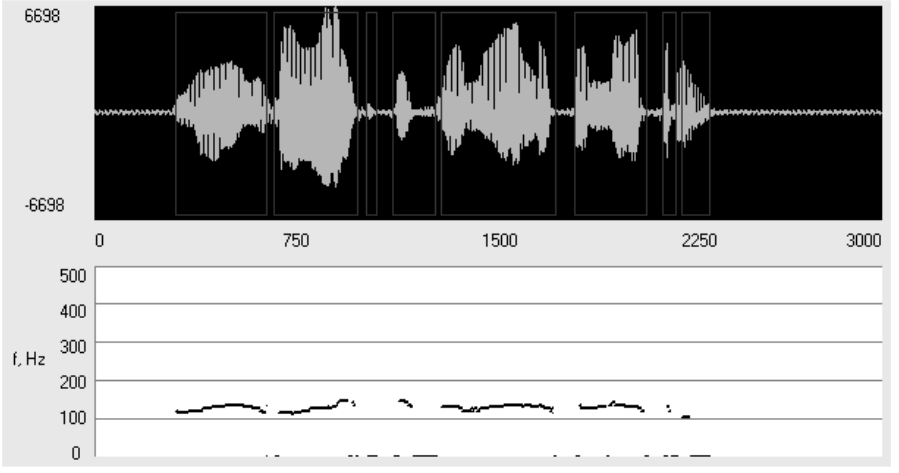
3.11 pav. Spektro koreliacinės funkcijos pavyzdys

Fig. 3.11. Example of the correlation function of spectrum

Viena iš problemų šiuo būdu skaičiuojant pagrindinį toną yra ta, kad kai kada pirmasis maksimumas arba nebūna išreikštas, arba būna labai silpnai išreikštas. Tuomet šis paieškos metodas suranda 2 ar net 3 harmoniką ir taip atsiranda žadinimo signalo pagrindinio dažnio skaičiavimo klaidos. Tam, kad būtų galima to išvengti, mes randame pirmąjį šios funkcijos maksimumą ir paimame du gretimus taškus, į vieną ir į kitą pusę nutolusius dvigubai mažesniu dažniu nei rasto maksimumo. Pvz., jei radome maksimumą ant 120 Hz, tai dar imame du šios funkcijos taškus ties 120±60 Hz, t. y. ant 60 ir 180 Hz. Toliau tikriname, ar rasto maksimumo amplitudės santykis su šių dviejų rastų taškų didesnis už tam tikrą slenkstį (pagal nutylėjimą slenksčio reikšmė lygi 1,5). Jei taip, tai tuomet imama ši žadinimo signalo pagrindinio dažnio reikšmė ir tikrinama, ar ji yra ribose nuo 70 iki 500 Hz. Jei taip, tas kadras laikomas vokalizuotu, priešingu atveju jis atmetamas. Tuo atveju, jei pirmoji harmonika nebūtų gerai išreikšta ir rastume maksimumą ant antros harmonikos, tai paėmę du gretimus taškus mes pataikytume ant pirmosios bei trečiosios harmonikų. Tuo atveju tikimybė, kad antroji harmonika bus gerokai labiau išreikšta už pirmą bei trečią, yra maža. Eksperimentiškai nustatyta, kad šis metodas labai retai daro klaidas, imdamas kitas kartotines žadinimo signalo pagrindinio dažnio harmonikas.

Vokalizuotų garsų išrinkimo algoritmo veikimo iliustravimas.

3.12 paveiksle iliustruotas sukurto vokalizuotų garsų išrinkimo (segmentavimo) algoritmo veikimas.



3.12 pav. Realizuoto vokalizuotų garsų išrinkimo algoritmo veikimo iliustravimas

Fig. 3.12. Illustration of the proposed algorithm of speech activity detection

Eksperimentas atliktas pasinaudojus sukurta programine įranga *gmm.exe*. Viršuje parodyta susegmentuota signalograma pagal algoritmus, aprašytus aukštesniuose punktuose, t. y. atmetus kadrus su labai mažomis signalo reikšmėmis, apskaičiavus foninio triukšmo slenkstį ir pašalinus pavienius impulsinius trukdžius. Susegmentuotos signalo dalys apvestos stačiakampiais. Apačioje parodyta žadinimo signalo pagrindinio dažnio kreivė.

Žadinimo signalo pagrindinio dažnio grafiko apačioje, kur šios reikšmės lygios 0, yra signalo vietos, kurios buvo paimtos aukštesniuose punktuose aprašytu segmentavimo algoritmu, bet jose nerastos tinkamos žadinimo signalo pagrindinio dažnio reikšmės. Šie kadrai taip pat atmetami ir toliau nebeanalizuojami. Galutiniam požymių išskyrimui naudojami tik tie kadrai, kuriuose žadinimo signalo pagrindinio dažnio reikšmės nelygios 0. Taip pat matome, kad šiame paveiksle nerasta nei vieno kadro, kur būtų paimta antroji žadinimo signalo pagrindinio dažnio harmonika.

3.2.4. Požymių išskyrimo sistemos kūrimas

Kalbančiojo atpažinimui naudojamos įvairios požymių sistemos, dalis kurių aptarta 2.3 skyriuje. Vieni iš naudojamų spektrinių požymių atitinka balso trakto parametrus, pvz. TPM parametrai, kepstras, apskaičiuotas iš TPM modelio (TPMK), formantės, spektrinės poros ir t. t. Asmens atpažinimui taip pat naudojami žadinimo signalo parametrai (žadinimo signalo pagrindinis dažnis F0, balsaskylės impulso formos modelis (Slyh *et al.* 2004)). Požymių sistemose, kurios skaičiuojamos iš Furjė spektro, atsispindi tiek balso trakto parametrai, tiek ir (dažnai netiesiogiai) žadinimo signalo (pvz. MSKK, barkų skalės kepstro koeficientai, kepstras, apskaičiuotas iš Furjė spektro ir t. t.).

Naudojamų požymių vektorių komponenčių skaičius būna įvairus. Pavyzdžiui naudojant TPM koeficientus, jų imama apie 12, naudojant TPMK, jų imama apie 18–22, naudojant MSKK, jų imama 13–15 ir t. t.

Kalbančiojo modeliavimui bei atpažinimui, naudojant Gauso mišinių modelius, tikslinant kalbėtojų modelių parametrus, reikia atlikti daug skaičiavimo operacijų ir šis procesas trunka palyginti ilgai. Skaičiavimo operacijų skaičius yra proporcingas naudojamų požymių vektorių komponenčių skaičiui, todėl patogiau naudoti kuo mažesnių išmatavimų požymių sistemą.

Mes siūlome apjungti žadinimo signalo ir balso trakto parametrus, ir kalbančiojo atpažinimui naudoti nedidelių išmatavimų, susidedančių iš 8 elementų, požymių vektorių sistemą, kurią sudaro: keturios formantės, trys antiformantės ir žadinimo signalo pagrindinis dažnis (F0).

Atpažinimo sistemoje realizuoti dviejų tipų požymiai:

- standartiniai melų skalės kepstro koeficientai (MSKK);
- pasiūlyta požymių sistema (formantės, antiformantės ir F0).

3.2.4.1. Melų skalės kepstro koeficientai

MSKK skaičiavimas aptartas 2.3.7 skyriuje, panaudojant (2.37)–(2.42) formules. Kiekvienam naudojamam kadrui skaičiuojama GFT, spektro taškų skaičius lygus 512. Po to suformuojamas trikampių filtrų rinkinys, išsidėstęs pagal melų skalę. Šių filtrų skaičius pagal nutylėjimą 25. Toliau skaičiuojamas melų skalės spektras ir pritaikius diskrečiąją kosinusų transformaciją randami melų skalės kepstro koeficientai. Pagal nutylėjimą atpažinimui naudojama pirmųjų 13 koeficientų.

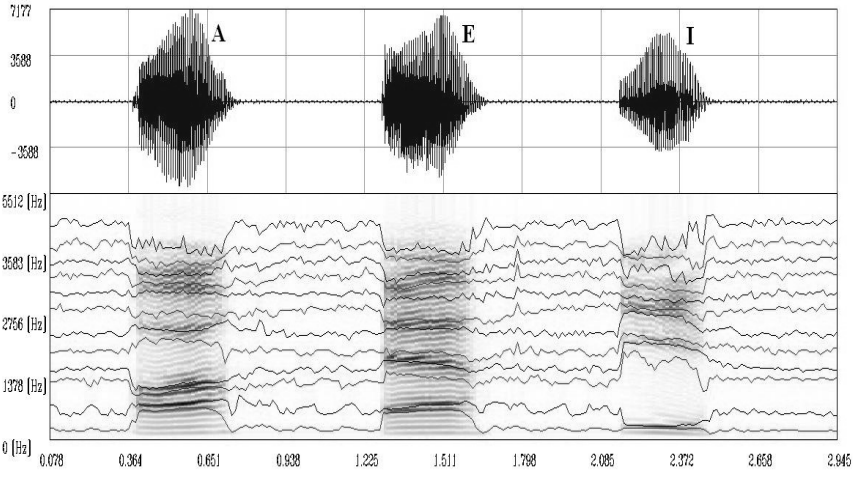
3.2.4.2. Žadinimo signalo ir balso trakto požymiai

Antra, mūsų pasiūlyta, atpažinimui skirtų požymių vektorių grupė susideda iš formančių (signalo spektro gaubtinės maksimumų dažnių), antiformančių (signalo spektro gaubtinės minimumų dažnių) ir žadinimo signalo pagrindinio dažnio reikšmių. Vartotojas pats gali laisvai pasirinkti norimą požymių vektoriaus elementų derinį. Pavyzdžiui, modelių kūrimui ir atpažinimui galima naudoti tik žadinimo signalo pagrindinį dažnį, tik formantes, tik antiformantes, arba požymius derinti, pvz. keturias formantes ir F0, tris formantes, dvi antiformantes ir F0 ir t. t. Didžiausia požymių vektoriaus dimensija, kai naudojami visi požymiai: keturios formantės, trys antiformantės ir žadinimo signalo pagrindinis dažnis – F0. Renkantis požymių vektorių sistemą, reiktų atsižvelgti į garso įrašų kokybę. Pvz. jei įrašytas kalbos signalas, turintis siaurą dažnių juostą (tarkime iki 2,5 kHz), geriau nenaudoti aukštesnių formančių, nes jos bus iškraipytos ir atpažinimo rezultatai gali pablogėti.

Žadinimo signalo pagrindinio dažnio radimas jau buvo aptartas aukščiau. Dabar aptarsime balso trakto požymių radimą.

Formantes galima rasti panaudojus (2.53) formulę ar iš TPM parametrų spektro gaubiamosios. Tačiau tai ne visuomet pavyksta. Esant prastesnei garso įrašo kokybei, TPM spektras tampa plokščias ir jame kai kurie maksimumai išnyksta. Mes formančių bei antiformančių įverčius randame iš spektrinių porų, panaudojus spektrinių porų metodą, aprašytą 2.3.10 punkte, kadangi jas apskaičiuoti visada galime. Pradžioje randamos spektrinės poros. Tam tikslui randami signalo kadro 12 eilės TPM koeficientai, sudaromas balso trakto perdavimo funkcijos polinomas bei jo veidrodinis atspindys, pagal (2.48) ir (2.49) formules. Toliau skaičiuojami suminis ir skirtuminis polinomai ((2.50) ir (2.51) formulės) ir randamos jų šaknys, kurios sudaro spektrines poras. 3.13 paveiksle pavaizduota signalograma, (įrašytos fonemos A E ir I) apačioje pavaizduota sonograma arba spektrograma (trumpalaikių signalo Furjė spektrų šeima, t. y. paimamas signalo kadras, apskaičiuojama jo Furjė transformacija ir

vertikaliai atidedama. Spektrogramos tamsumas rodo spektro intensyvumą). Spektrogramoje galima įžvelgti tam tikras tamsesnes juostas, t. y. formantės.



3.13 pav. Signalograma, jos sonograma ir spektrinių porų pavaizdavimas

Fig. 3.13. Signalogramm, spectrogram and linear spectral pairs

Matome, kad A ir E fonemos turi gana ryškiai išreikštas keturias formantes. Šiame paveiksle taip pat kreivėmis pavaizduotos spektrinės poros, apskaičiuotos ankščiau paminėtu algoritmu. Kaip matome iš šio paveikslo, spektrinės poros apytiksliai apgaubia formantes, t. y. vieną iš spektrinių porų dažnių galima priskirti formantei. Taip pat mes darome prielaidą, kad antiformantės išsidėsto tarp dviejų gretimų formančių.

Tegul $LSF(N)$ žymi N-tąją spektrinių porų dažnių reikšmę, $F(M)$ – M-tąją formantę, $ANF(K)$ – K-tąją antiformantę. Mes naudojame tokį formančių ir antiformančių skaičiavimą (Salna, Mambro 2006):

$$\begin{cases} F(1) = LSF(2); \\ F(2) = LSF(5); \\ F(3) = LSF(8); \\ F(4) = LSF(11); \\ ANF(1) = (LSF(2)+LSF(3))/2; \\ ANF(2) = (LSF(5)+LSF(6))/2; \\ ANF(3) = (LSF(8)+LSF(9))/2. \end{cases} \tag{3.10}$$

T. y. pirmoji formantė prilyginama antrąjam spektrinių porų dažniui, pirmoji antiformantė – antros ir trečios spektrinių porų dažnių aritmetiniam vidurkiui ir t. t.

Kadangi aukštesnių formančių dispersija yra didesnė nei žemesnių, tam, kad būtų galima „suvienodinti“ visų formančių ir antiformančių dispersijas mes siūlome formantes bei antiformantes skaičiuoti melų skalėje, kuri iki 1 kHz yra maždaug tiesinė, toliau logaritminė.

3.2.5. Kalbančiųjų modelių kūrimas

Kalbančiojo balso savybių modeliavimui panaudoti standartiniai Gauso mišinių modeliai, aprašyti (2.65)–(2.67) išraiškose. Jeigu naudojami požymiai yra tarp savęs nepriklausomi (nekoreliuoti), tuomet išraiškoje (2.66) kovariacinė matrica tampa diagonalinė ir šią išraišką galime supaprastintai užrašyti:

$$b_i(\vec{x}) = \prod_{i=1}^{D} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x_i - \mu_i)^2}{2} \right). \tag{3.11}$$

Nekoreliuotais požymiais galima laikyti kepstro koeficientus, formantes.

3.2.5.1. Pradinis GMM parametrų vertinimas

Viena iš problemų kuriant Gauso mišinių modelius yra pradinis modelio parametrų vertinimas (Reynolds, Rose 1995). Dažniausiai parenkamos atsitiktinės parametrų vertės, kurios po to su kiekviena iteracija tikslinamos. Arba pradinė požymių vektorių erdvė padalijama į tam tikrą skaičių sričių (klasterių), po to randami reikiami šių klasterių statistiniai parametrai ir priskiriami atitinkami mišinio komponentei. Sukurtoje kalbančiojo atpažinimo sistemoje taip pat realizuotas pradinių parametrų vertinimas dalijant pradinę požymių vektorių aibę į nepersikertančius klasterius, kurių skaičius lygus Gauso mišinių komponenčių skaičiui. Tuomet skaičiuojamos kiekvieno klasterio statistinės charakteristikos (vidurkis, standartinė deviacija) ir priskiriamos atitinkamam Gauso mišiniui kaip pradiniai parametrai. Mišinio pradinis svoris randamas kaip požymių vektorių, sudarančių klasterį, skaičiaus santykis su visų požymių vektorių skaičiumi. Tegul T_i – *i*-tojo klasterio dydis (jį sudarančių požymių vektorių skaičius), T – bendras visų požymių vektorių skaičius, $\vec{x}_{i,t}$ – *t*-tasis požymių vektorius, esantis klasteryje *i*. Tuomet pradiniai Gauso mišinio *i*-tosios komponentės parametrai:

$$\vec{\mu}_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\vec{x}_{i,t}\,, \tag{3.12}$$

$$\vec{\sigma}_i = \sqrt{\frac{1}{T_i}\sum_{t=1}^{T_i}(\vec{x}_{i,t}-\vec{\mu}_i)^2}\,, \tag{3.13}$$

$$p_i = \frac{T_i}{T}\,. \tag{3.14}$$

Šioje sistemoje realizuoti trys klasterių formavimo metodai:

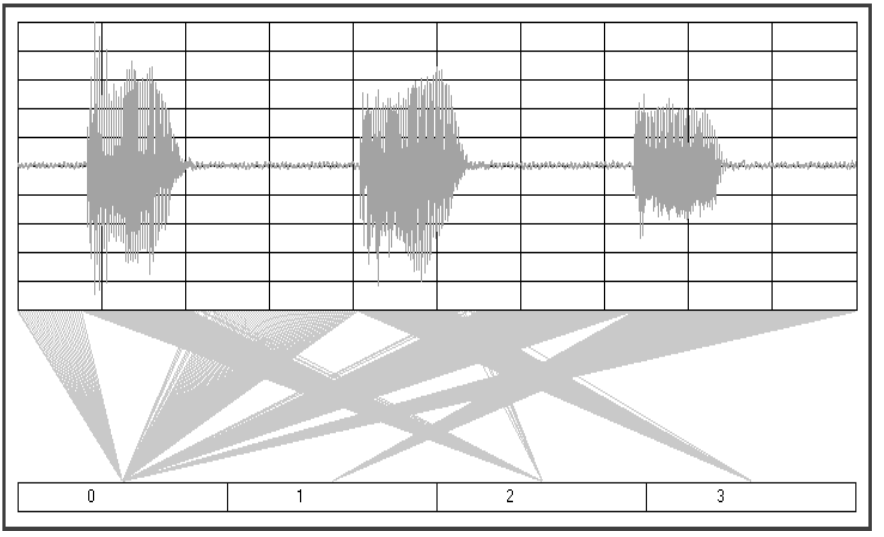
- tiesinis dalijimas į klasterius;
- atsitiktinis klasterių formavimas;
- vektorinio kvantavimo (VK) metodas.

Tiesinio dalijimo į klasterius atveju visi vokalizuotų kadrų požymių vektoriai nuosekliai sudalijami į lygias dalis ir sudaro atitinkamus klasterius.

Atsitiktinio klasterių formavimo atveju kiekvienas požymių vektorius atsitiktinai priskiriamas vienam iš klasterių.

Kadangi, kuriant Gauso mišinių modelius, matematinės vilties maksimizavimo algoritmas pasiekia lokalius maksimumus, galutinis modelio tikėtinumas gali priklausyti nuo pradinių parametrų įverčio. Dėl to pradinių parametrų vertinimo būdas gali įtakoti atpažinimo rezultatus.

Mes siūlome, vertinant pradinius GMM parametrus, klasterių formavimą atlikti pasinaudojant modifikuotu *LBG vektorinio kvantavimo* algoritmu (Lipeika, Lipeikienė 1995), kuomet kiekvieno dalijimo metu klasterių skaičius didinamas vienetu, „iškraipant“ centroidą, turintį didžiausius iškraipymus. Dalijant centroidą į du, „iškraipomi“ atspindžio koeficientai. Kaip žinoma, taikant vektorinį kvantavimą pradinė požymių vektorių erdvė suskaidoma į pasirinktą skaičių klasterių. Į klasterius sugrupuojami panašių garsų požymių vektoriai. 3.14 paveiksle pavaizduotas kalbos signalas, kur buvo ištartos raidės „A“, „E“ ir „I“. Parinkus, pvz. keturis klasterius, kaip matome iš šio paveikslo (čia panaudota vektorinio kvantavimo programinė įranga ***VeckeyPr.exe***, grafiškai vaizduojanti klasterizacijos procesą), atlikus klasterizaciją, nuliniam klasteriui priskiriamos signalo dalys, atitinkančios triukšmus. Pirmą klasterį atitinka fonemą „I“, antrajam klasteriui daugiausiai priskirta požymių vektorių, atitinkančių fonemą „A“. Trečiajam klasteriui priskirti požymių vektoriai, atitinkantys fonemą „E“.



3.14 pav. Signalo kadrų priskyrimo keturiems klasteriams pavyzdys

Fig. 3.14. Assigning of signal frames to the four clusters

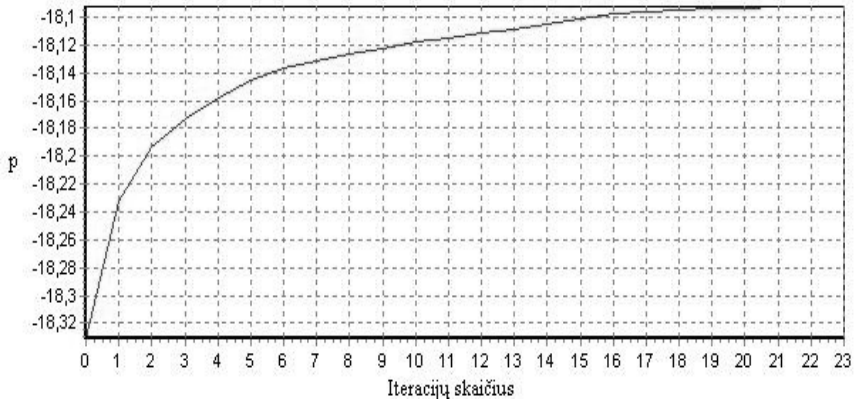
Taigi, sugrupavus į atskirus klasterius panašius garsus, galima kiekvienos iš požymių vektoriaus komponenčių pasiskirstymą klasteryje aprašyti atskiru Gauso pasiskirstymu. Dėl šių priežasčių, taikant vektorinį kvantavimą, pagerėja pradinių GMM parametrų įvertis.

3.2.5.2. GMM parametrų vertinimas

Vertinant GMM parametrus naudojamas matematinės vilties maksimizavimo (MVM) algoritmas. Tuo tikslu kiekvienos iteracijos metu pagal (2.70)–(2.73) formules perskaičiuojami GMM parametrai. Perskaičiavus GMM parametrus, toliau pagal (2.69) formulę skaičiuojamas GMM tikėtinumas, tik skaičiuojant tikėtinumų logaritmus ir normuojant iš požymių vektorių skaičiaus T:

$$p(X \mid \lambda) = \frac{1}{T} \sum_{t=1}^{T} \lg(p(\vec{x}_t \mid \lambda)) . \tag{3.15}$$

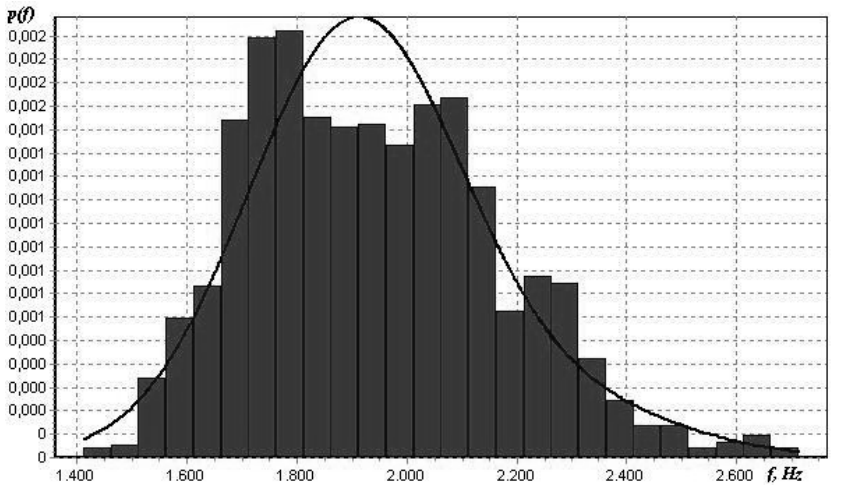
Šių parametrų tikslinimas nutraukiamas, kai po iteracijos modelio tikėtinumas padidėja ne daugiau kaip 0,001 % lyginant su prieš tai buvusia reikšme (šį parametrą galima pakeisti).



3.15 pav. GMM tikėtinumo augimo pavyzdys

Fig. 3.15. Example of increasing likehood of GMM

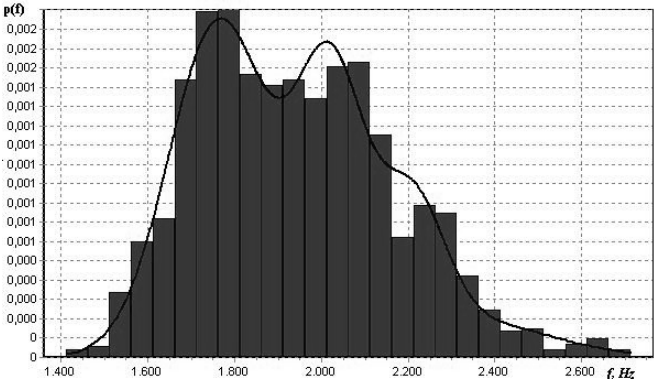
GMM tikėtinumo augimo pavyzdys pateiktas 3.15 paveiksle. Čia įvykdytos 23 iteracijos. Požymiais panaudota keturios formantės, trys antiformantės ir F0. Požymių vektorių skaičius lygus 1 474 ir panaudotas Gauso mišinys, sudarytas iš 15 komponenčių (pasvertų Gauso funkcijų).



3.16 pav. Antrosios formantės (melų skalėje) histograma ir jos aproksimavimas GMM kai paimtos 2 komponentės

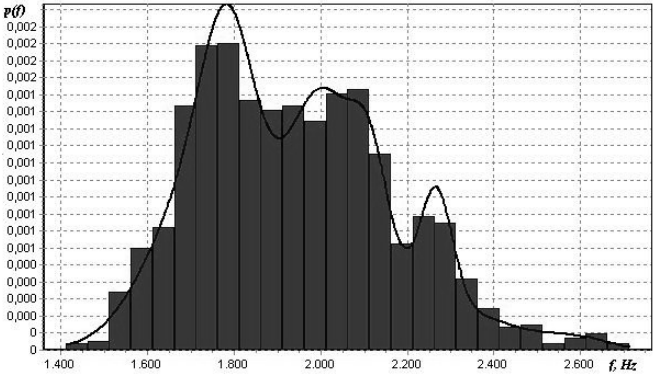
Fig. 3.16. Histogram of second formant (in mel scale) and approximation by GMM when 2 components are used

3.16 paveiksle pavaizduota antrosios formantės tikrojo pasiskirstymo ir gauto aproksimavimo Gauso mišinių modeliu, kai paimtos 2 GMM komponentės, pavyzdys. 3.17 paveiksle, kai paimtos 5 GMM komponentės, ir 3.18 paveiksle, kai paimta 15 GMM komponenčių. Požymių vektorių skaičius lygus 1 474, naudojamos 4 formantės, trys antiformantės ir F0.



3.17 pav. Antrosios formantės (melų skalėje) histograma ir jos aproksimavimas GMM kai paimtos 5 komponentės

Fig. 3.17. Histogram of second formant (in mel scale) and approximation by GMM when 5 components are used



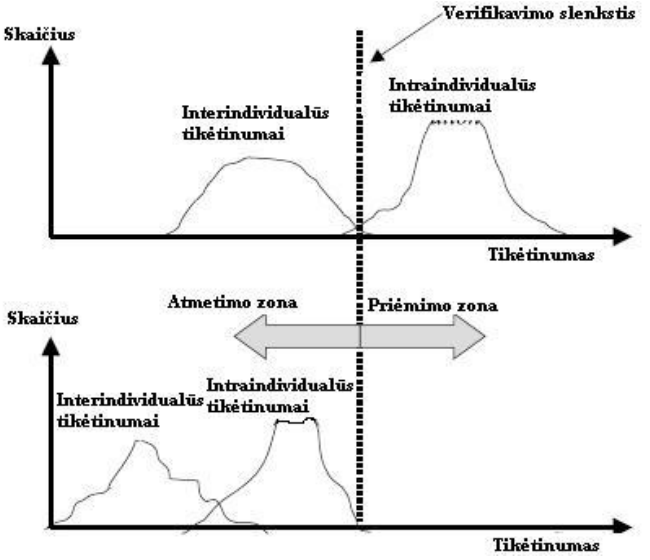
3.18 pav. Antrosios formantės (melų skalėje) histograma ir jos aproksimavimas GMM kai paimta 15 komponenčių

Fig. 3.18. Histogram of second formant (in mel scale) and approximation by GMM when 15 components are used

3.2.6. Mokymas ir slenksčio nustatymas

Mokymo tikslas yra kiekvienam kalbėtojui nustatyti slenkstį, naudojamą verifikavimui ar atviros aibės identifikavimui. Mokymo metu, analogiškai, kaip ir kalbėtojo modelio kūrimo metu, nuskaitomas kalbos signalas, apdorojamas, segmentuojamas ir skaičiuojami požymių vektoriai. Toliau tie požymių vektoriai lyginami su kiekvieno kalbėtojo modeliu. Lyginant skaičiuojamas tikėtinumas pagal (3.15) formulę, kurioje normavimas iš vektorių skaičiaus reikalingas tolesniam slenksčio nustatymui tam, kad suvienodintume tikėtinumą, nes nenormuojant, jis labai priklausytų nuo požymių vektorių skaičiaus. Jei lyginama frazė su to paties asmens modeliu, gautas tikėtinumas pridedamas prie to asmens intraindividualių tikėtinumų masyvo, priešingu atveju – prie interindividualių. Taigi, kiekvienam sistemoje registruotam kalbėtojui palyginus jo pasakytas frazes su jo ir kitų kalbėtojų modeliais, kiekvienam iš kalbėtojų gaunami intraindividualių ir interindividualių tikėtinumų masyvai.

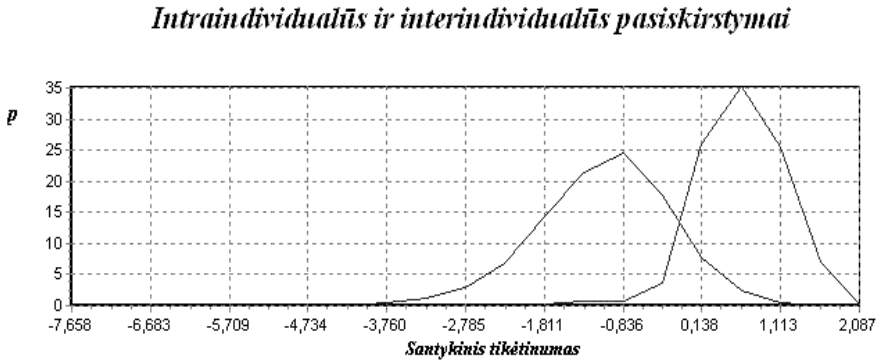
Viena iš problemų yra ta, kad kiekvienam kalbėtojui šie tikėtinumai yra labai skirtingi. Tai iliustruota 3.19 paveiksle. Skirtingų kalbėtojų šios kreivės būna gerokai pasislinkusios viena kitos atžvilgiu.



3.19 pav. Dviejų kalbėtojų intraindividualių ir interindividualių panašumo įverčių pasiskirstymo pavyzdys

Fig. 3.19. Intraindividual and interindividual similarity scores of two speakers

Taigi, nėra galimybės nustatyti visiems kalbėtojams bendro vienodo slenksčio, dėl to slenkstis užduodamas individualus. Sistemoje yra du būdai, kaip nustatyti slenkstį. Vienas būdas, kiekvienam asmeniui galima šį slenkstį nustatyti tiesiogiai, stebint to asmens intraindividualių ir interindividualių tikėtinumų pasiskirstymų kreives. Antruoju būdu yra skaičiuojama bendra normuota intraindividualių ir interindividualių tikėtinumų pasiskirstymo kreivė. Tam tikslui pradžioje skaičiuojamos intraindividualių ir interindividualių tikėtinumų pasiskirstymo kreivės kiekvienam asmeniui, randamas individualus LKL taškas, kuris tampa normavimo koeficientu. Vėliau, kiekviena iš kreivių normuojama, t. y. iš kiekvienos intraindividualių ir interindividualių tikėtinumų reikšmės atimama tikėtinumo vertė, atitinkanti šį LKL tašką. Tuomet visos kreivės pasistumia ir visų jų LKL taškai atsiduria ant nulinės reikšmės. Tada jau galima brėžti bendrą normuotą intraindividualių ir interindividualių tikėtinumų pasiskirstymų kreivę. Šios kreivės pavyzdys pateiktas 3.20 paveiksle. LKL taškas randamas keičiant slenksčius (tikėtinumus) ir skaičiuojant „savojo“ atmetimo (KA) ir „svetimo“ priėmimo (KP) tikimybes. Tikėtinumas, prie kurio šių tikimybių skirtumas lygus 0 arba minimalus, atitiks LKL tašką. „Savojo“ atmetimo tikimybė prie nustatyto slenksčio bus lygi santykiui skaičiaus intraindividualių taškų, kurių reikšmės mažesnės už nustatytą slenkstį su visų intraindividualių taškų skaičiumi. „Svetimo“ priėmimo tikimybė prie nustatyto slenksčio bus lygi santykiui skaičiaus interindividualių taškų, kurių reikšmės didesnės už nustatytą slenkstį su visų interindividualių taškų skaičiumi.

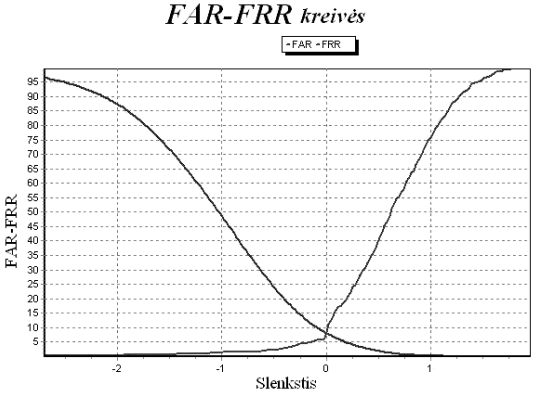


3.20 pav. Bendra intraindividualių ir interindividualių tikėtinumų pasiskirstymų kreivė

Fig. 3.20. Common curve of distribution of intraindividual and interindividual likehoods

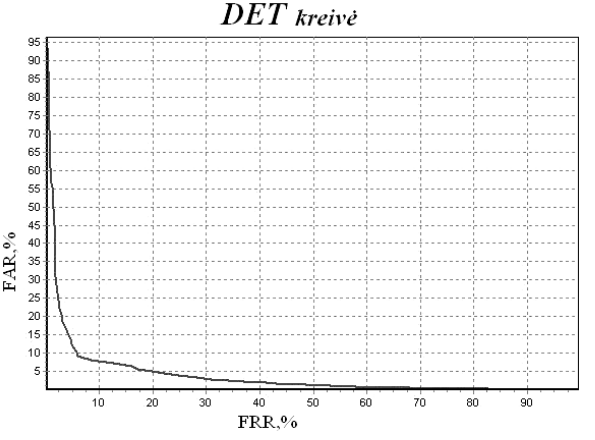
Nustačius bendrą normuotą slenkstį (jei tai bus LKL taškas, tuomet nurodomo slenksčio reikšmė bus 0), toliau sistema apskaičiuoja individualų slenkstį kiekvienam kalbėtojui. Šis slenkstis apskaičiuojamas prie įvesto normuoto slenksčio pridėjus to asmens normavimo koeficientą, t. y. atstatomas to asmens tikrasis LKL taškas.

Įvykdžius mokymą, galima pažiūrėti sistemos veikimo tikslumo rezultatus: KPL-KAL (3.21 paveikslas) ir DET kreives (3.22 paveikslas).



3.21 pav. KPL-KAL kreivių pateikimas

Fig. 3.21. FAR-FRR curves



3.22 pav. DET kreivės pateikimas

Fig. 3.22. DET curve

3.2.7. Atpažinimas

Kaip jau buvo minėta, sistemoje realizuoti visi asmens atpažinimo pagal balsą algoritmų tipai: atviros ir uždaros aibės kalbančiojo identifikavimas ir verifikavimas. Atviros aibės kalbančiojo identifikavime ir verifikavime turi būti įvykdytas mokymas ir nustatytas slenkstis. Atpažinimo metu analogiškai atidaroma garso rinkmena, nuskaitomas signalas, apdorojamas, segmentuojamas, skaičiuojami požymių vektoriai. Jei pasirinktas uždaros aibės identifikavimas, sistema skaičiuoja kiekvieno kalbėtojo modelio tikėtinumą duotai požymių vektorių sekai pagal (3.14) formulę. Asmuo, kurio modeliui gaunamas didžiausias tikėtinumas, grąžinamas kaip atpažinimo rezultatas. Atviros aibės kalbančiojo identifikavime ir verifikavime šis tikėtinumas dar lyginamas su slenksčiu, jei jis viršija slenkstį, asmuo atpažintas.

3.3. Trečiojo skyriaus apibendrinimas

- Sukurta nepriklausoma nuo ištartos frazės kalbančiojo atpažinimo sistema.
- Pasiūlytas automatinis vokalizuotų garsų išskyrimo metodas. Šis metodas nereikalauja iš vartotojo jokių papildomų veiksmų, tokių kaip kalbos signalo ir triukšmo pavyzdžių nurodymo ir t. t.
- Pasiūlyta nedidelio komponenčių skaičiaus požymių vektorių sistema, susidedanti iš žadinimo signalo parametrų ir balso trakto parametrų. Kaip žadinimo signalo parametras panaudotas žadinimo signalo pagrindinis dažnis, kaip balso trakto parametrai, panaudotos keturios formantės (signalo spektro gaubtinės maksimumų dažniai) bei trys antiformantės (signalo spektro gaubtinės minimumų dažniai).
- Formančių ir antiformančių radimui panaudotas spektrinių porų metodas, kadangi tiesiogiai jas rasti ne visada galima.
- Siekiant suvienodinti visų formančių bei antiformančių dispersijas, pasiūlyta jas skaičiuoti melų skalėje.
- Pradinių GMM parametrų vertinimas realizuotas panaudojant klasterizacijos principą. Klasterių statistiniai parametrai panaudojami kaip pradiniai GMM parametrai.
- Vertinant pradinius GMM parametrus pasiūlyta klasterių formavimui panaudoti modifikuotą vektorinio kvantavimo algoritmą.

4

Atpažinimo sistemos eksperimentinis tyrimas

Šiame skyriuje bus pateikti atpažinimo sistemos eksperimentinio tyrimo rezultatai. Pagrindinis eksperimentų tikslas – ištirti ir palyginti pasiūlytų požymių vektorių, susidedančių iš žadinimo signalo ir balso trakto parametrų (formančių ir antiformančių) su vienais iš plačiausiai naudojamų pasaulyje – melų skalės kepstro koeficientų, atpažinimo tikslumą, esant vienodoms visoms kitoms sąlygoms. Taip pat bus tiriama ir lyginama atpažinimo tikslumo priklausomybė nuo skirtingo Gauso funkcijų, sudarančių Gauso mišinius (komponenčių), skaičiaus bei kitų parametrų, tokių kaip formančių skaičiavimas hercų ir melų skalėje, o taip pat nuo pradinio GMM parametrų vertinimo metodo.

Skyriaus tematika paskelbtas autoriaus straipsnis (Kamarauskas 2008).

4.1. Eksperimentų sąlygos ir duomenys

Eksperimentai atlikti sukurta kalbančiojo asmens atpažinimo sistema *gmm.exe*, panaudojant asmeninį kompiuterį. Kalbos signalai įvedami iš garsinių rinkmenų, panaudojant balsų bazę. Kadangi pagrindinis tyrimų tikslas yra mūsų pasiūlytų ir vienų populiariausių pasaulyje požymių vektorių palyginimas, klausimas dėl signalų apdorojimo parametrų optimalumo nekeltas. Tokie signalų

apdorojimo parametrai, kaip analizės kadro ilgis ir jo postūmis, pradinės filtracijos koeficientas, tiesinės prognozės eilė ir t. t., nebuvo keičiami ir buvo parinkti remiantis autoriaus patirtimi.

Atpažinimo rezultatai pateikiami grafikuose. Pateiktos kiekvieno eksperimento metu gautos DET kreivės, nurodyta LKL taško reikšmė bei kai kurios intraindividualių ir interindividualių pasiskirstymų kreivės.

Eksperimentai atlikti panaudojant STC (Speech Technology Center) balsų bazę, įrašytą 1996–1998 metais. Pagrindinis šios balsų bazės kūrimo tikslas kalbančiojo atpažinimo algoritmų tikrinimas bei kalbėtojų individualaus kintamumo tyrimai. Balsų bazė įrašyta panaudojant 16 bitų Vibra-16 Creative Labs garso plokštę bei 11 025 Hz diskretizacijos dažnį. Toliau trumpai pateiksime detalesnį šios balsų bazės aprašymą.

Kalba. Rusų.

Kalbėtojai. Balsų bazėje įrašyta 89 skirtingų kalbėtojų (iš jų 54 vyrai ir 35 moterys) ištartos frazės. Iš jų 69 kalbėtojai (36 vyrai ir 33 moterys) dalyvavo 15 ir daugiau įrašymo sesijų, 11 kalbėtojų su 10 ar daugiau sesijų ir 9 kalbėtojai su mažiau kaip 10 įrašymo sesijų. Kalbėtojų gimtoji kalba rusų, amžius svyruoja tarp 18–62 metų, įrašai daryti Sankt-Peterburge.

Rinkinys. Rinkinys susideda iš 5 sakinių. Kiekvienas kalbėtojas atsargiai, bet laisvai ir sklandžiai 15 kartų perskaito kiekvieną sakinį skirtingu laiku, 1–3 mėnesių laikotarpyje.

Rinkinys susideda iš 6 856 frazių, įrašytų dviejose kompaktinėse plokštelėse, bendras dydis apie 837 MB. Kiekvienos ištartos frazės signalas saugojamas atskiroje garso rinkmenoje (kurios dydis apytiksliai 126 KB). Vidutinė ištartos frazės trukmė apie 5 sek. Tam, kad būtų galima nustatyti mikrofono kokybės įtaką, dalis balsų bazės įrašyta panaudojant prastos kokybės mikrofoną.

Įrašymo sąlygos. Panaudotas dinamiškas, įvairiakryptis, aukštos kokybės mikrofonas, atstumas iki burnos 5–15 cm. Įrašymo aplinka – ofiso kambarys. Diskretizacijos dažnis 11 025 Hz, rezoliucija – 16 bitų, garso plokštė: Creative Labs Vibra-16.

Ši balsų bazė platinama ELRA (European Language Resources Association).

Mes eksperimentams atrinkome dalį šios balsų bazės. Buvo paimta 41 vyro ištarta pirma frazė „*Ne ishchite stenografistku yesli u vas uzhe yest transkraiber i zhelaniye rabotat*“. Įrašai daryti geros kokybės mikrofonu. Dažniausiai įrašyta ta pati frazė po 15 kartų, kai kurių kalbėtojų įrašyta po 6 kartus, maždaug trečdalis įrašė apie 20 kartų. Kalbėtojai, kurių įrašų kiekis neviršija 6, nebuvo įtraukti tyrimams. Pirmosios trys kiekvieno kalbėtojo frazės buvo panaudotos to kalbėtojo Gauso mišinių modelio kūrimui, kitos atpažinimui.

4.2. Kalbančiojo atpažinimo tyrimai

Eksperimentų metu buvo tiriama pasiūlyta požymių vektorių sistema, susidedanti iš žadinimo signalo pagrindinio dažnio, formančių ir antiformančių. Buvo tiriami įvairūs pasiūlytos požymių sistemos elementų deriniai:

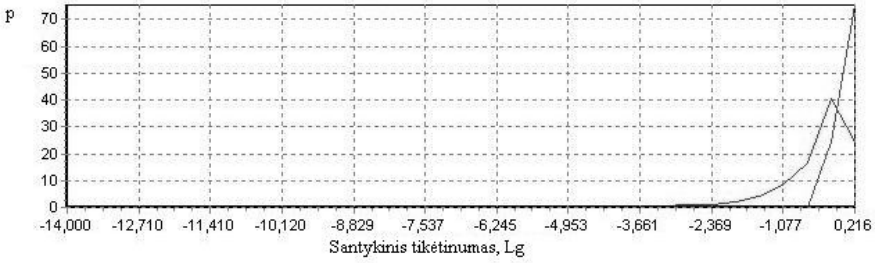
- naudojamas tik žadinimo signalo pagrindinis dažnis F0;
- naudojami tik balso trakto parametrai: keturios formantės;
- naudojami tik balso trakto parametrai: keturios formantės kartu su trimis antiformantėmis;
- naudojamas žadinimo signalo pagrindinis dažnis kartu su keturiomis formantėmis;
- naudojamas žadinimo signalo pagrindinis dažnis kartu su keturiomis formantėmis ir trimis antiformantėmis;

Eksperimentai buvo atliekami ir su standartiniais 13 eilės melų skalės kepstro koeficientais (MSKK).

Taip pat buvo tiriama ir atpažinimo tikslumo priklausomybė nuo skirtingo Gauso funkcijų (komponenčių), sudarančių Gauso mišinius, skaičiaus. Eksperimentų metu buvo naudojami Gauso mišiniai, sudaryti iš 5, 10, 15, 20 komponenčių bei adaptyvaus komponenčių skaičiaus, kuomet jis automatiškai parenkamas atsižvelgiant į modelio kūrimui skirtų požymių vektorių skaičių. Šiuo atveju mes atlikome eksperimentus naudodami Gauso mišinius su komponenčių skaičiumi, lygiu požymių vektorių skaičiui padalintam iš 70. Kadangi visos kitos eksperimentų sąlygos visiškai vienodos, t. y. visais atvejais imami tie patys signalų kadrai, nes naudojamas tas pats vokalizuotų garsų išskyrimo algoritmas, pagal gautus rezultatus galima atlikti požymių vektorių sistemų efektyvumo palyginimą.

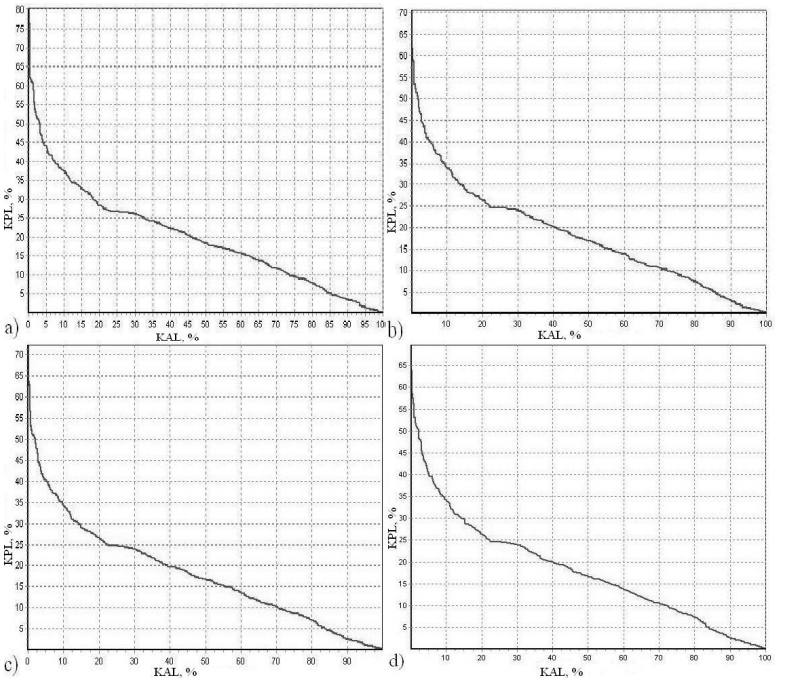
4.2.1. Kalbančiojo atpažinimas panaudojant žadinimo signalo pagrindinį dažnį

Pradžioje buvo atliekami tyrimai kalbos signalo požymiais naudojant žadinimo signalo pagrindinį dažnį (F0). Tyrimai buvo atlikti panaudojant Gauso mišinius, sudarytus iš 1, 3, 5 ir 10 komponenčių (pasvertų Gauso funkcijų) sumos. Mokymo metu gautos intraindividualių ir interindividualių tikėtinumų pasiskirstymų kreivės, kai panaudoti Gauso mišiniai, sudaryti iš 3 komponenčių, pavaizduotos 4.1 paveiksle.



4.1 pav. Interindividualių ir intraindividualių tikėtinumų kreivės kai panaudoti požymiai yra žadinimo signalo pagrindinis dažnis ir parinkti Gauso mišiniai, sudaryti iš 3 komponenčių

Fig. 4.1. Curves of intraindividual and interindividual likehoods when pitch was used as a feature and GMM consist of 3 components



4.2 pav. DET kreivės, gautos požymiais naudojant žadinimo signalo pagrindinį dažnį F0 ir panaudojant GMM komponenčių: a) vieną; b) tris; c) penkias; d) dešimt

Fig. 4.2. DET curves when pitch was used as a feature and count of GMM components: a) one; b) three; c) five; d) ten

4.2 paveiksle pateiktos klaidingo „svetimo“ priėmimo tikimybės (KPL) priklausomybės nuo klaidingo „savojo“ atmetimo tikimybės (KAL) kreivės – DET, kai kalbėtojo modeliui aprašyti panaudota 1, 3, 5 ir 10 Gauso mišinių komponenčių.

Iš 4.2 paveikslo a) matome, kuomet žadinimo signalo pagrindiniam dažniui aprašyti naudojama viena Gauso funkcija, gautojo lygių klaidų lygio reikšmė LKL=26,65 %. Iš šios kreivės taip pat matome, kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 80,5 %. Taip pat matome, kai klaidingo „svetimo“ priėmimo tikimybė lygi 0 (KPL=0 %), tuomet KAL=100 %.

4.2 paveiksle b) pavaizduota DET kreivė, kuomet žadinimo signalo pagrindiniam dažniui aprašyti naudojamos trys Gauso mišinių komponentės. Šiuo atveju gautojo lygių klaidų lygio reikšmė yra LKL=24,62 %. Iš šios kreivės taip pat matome, kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 70,6 %. Taip pat matome, kai KPL=0 %, tuomet KAL=100 %.

4.2 paveiksle c) pavaizduota DET kreivė, kuomet žadinimo signalo pagrindiniam dažniui aprašyti naudojamos penkios GMM komponentės. Šiuo atveju gauta LKL (lygių klaidų lygio) taško reikšmė yra lygi 24,7 %. Kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 72,5 %. Taip pat matome, šiuo atveju, kai klaidingo „svetimo“ priėmimo tikimybė lygi 0 (KPL=0 %), tuomet KAL=100 %.

4.2 paveiksle d) pateikta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 10 GMM komponenčių. Šiuo atveju gautojo LKL taško reikšmė yra 24,6 %. Iš šių kreivių taip pat matome, kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 69,8 %. Taip pat matome, kai klaidingo „svetimo“ priėmimo tikimybė lygi 0 (KPL=0 %), tuomet KAL=100 %.

4.1 lentelė. Atpažinimo rezultatai panaudojant žadinimo signalo pagrindinį dažnį

Table 4.1. Recognition results when pitch was used

GMM komponenčių skaičius	LKL, %	KPL0, % (KAL vertė kai KPL=0)	KAL0, % (KPL vertė kai KAL=0)
1	26,65	100	80,5
3	24,62	100	70,6
5	24,7	100	72,5
10	24,6	100	69,8

Atpažinimo rezultatai taip pat pateikti 4.1 lentelėje, kur nurodoma LKL (lygių klaidų lygio) vertė, taip pat tikimybė, kad „savas“ bus atmestas kaip „svetimas“, (pažymėta KPL0), kuomet „svetimo“ priėmimo tikimybė lygi 0, t. y. KPL=0 % bei tikimybė, kad „svetimas“ bus priimtas kaip „savas“, (pažymėta KAL0), kuomet „savojo“ atmetimo tikimybė lygi 0, t. y. KAL=0 %.

Analizuojant gautus rezultatus matome, kad panaudojant požymiais tik žadinimo signalo pagrindinį dažnį, gautas gana didelis klaidų lygis, apie 25 % ir mažai priklauso nuo Gauso funkcijų, sudarančių mišinius, skaičiaus. Didelį klaidų lygį gauname dėl to, kad esant trumpoms frazėms vien tik šis požymis nėra labai tinkamas, kadangi kiekvieną kartą ištarti trumpą frazę su ta pačia intonacija yra sudėtinga, ir tai tiesiogiai įtakoja pagrindinį toną. Be to, žadinimo signalo pagrindinis dažnis apytiksliai pasiskirsto pagal Gauso dėsnį, dėl šios priežasties jo pasiskirstymui aprašyti nereikia daug Gauso mišinių komponenčių. Iš atliktų eksperimentų galima daryti išvadą, kad žadinimo signalo pagrindiniam dažniui aprašyti pilnai pakanka mišinio, sudaryto iš trijų Gauso funkcijų. Atpažinimo tikslumas panaudojant vieną Gauso funkciją ir mišinius, sudarytus iš 10 Gauso funkcijų skiriasi tik 2,05 %.

Atlikti eksperimentai rodo, kad žadinimo signalo pagrindinis dažnis nėra geras ir patikimas požymis kalbančiojo atpažinimui, tačiau jo privalumas – gana didelis atsparumas įvairiems įrašymo kanalo iškraipymams bei iškraipymams, atsirandantiems dėl įrašymo sąlygų neatitikimo.

4.2.2. Kalbančiojo atpažinimas panaudojant balso trakto parametrus

Sekantis tyrimų etapas, požymių, atitinkančių balso traktą, tyrimas. Iš šių požymių eliminuota žadinimo signalo įtaka. Eksperimentai atlikti panaudojant Gauso mišinius, sudarytus iš 5, 10, 15, 20 bei adaptyvaus komponenčių skaičiaus. Požymiais buvo panaudotos keturios formantės bei keturios formantės kartu su trimis antiformantėmis. Formantės ir antiformantės buvo skaičiuotos melų skalėje.

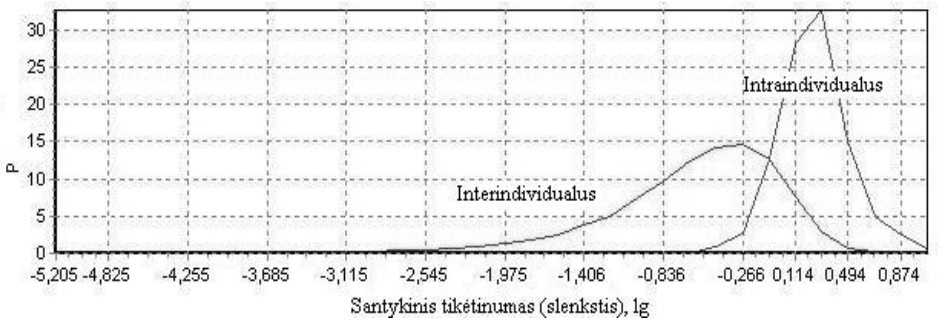
4.2.2.1. Kalbančiojo atpažinimas panaudojant keturias formantes melų skalėje

Mokymo metu gautos intraindividualių ir interindividualių tikėtinumų pasiskirstymų kreivės, kai panaudoti Gauso mišiniai, sudaryti iš 20 Gauso funkcijų (komponenčių) sumos, pavaizduotos 4.3 paveiksle.

4.4 paveiksle pateiktos klaidingo „svetimo“ priėmimo tikimybės (KPL) priklausomybės nuo klaidingo „savojo“ atmetimo tikimybės (KAL) kreivės – DET, kai panaudoti Gauso mišiniai, sudaryti iš 5, 10, 15, 20 ir adaptyvaus

skaičiaus Gauso funkcijų (komponenčių) sumos ir požymiais naudojant keturias formantes melų skalėje.

4.4 paveiksle a) pateikta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 5 Gauso funkcijų sumos. Gautoji LKL (lygių klaidų lygio) taško reikšmė yra 13,8 %. Iš šios kreivės taip pat matome, kuomet klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 54,8 %. Taip pat matome, kai klaidingo „svetimo“ priėmimo tikimybė lygi 0 (KPL=0 %), tuomet KAL=98 %.

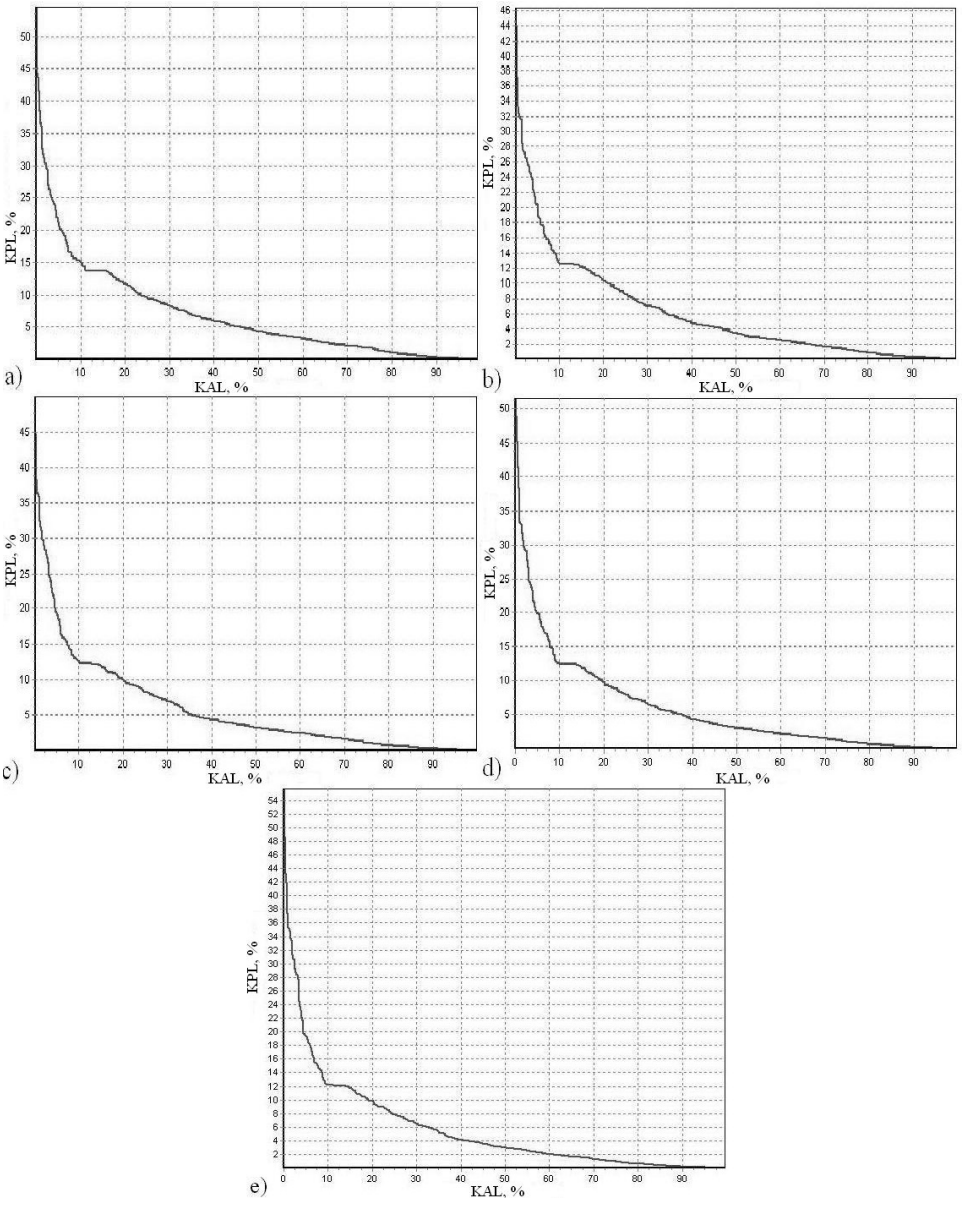


4.3 pav. Interindividualių ir intraindividualių tikėtinumų pasiskirstymų kreivės kai panaudoti požymiai yra keturios formantės melų skalėje ir parinkta Gauso mišiniai, sudaryti iš 20 GMM komponenčių

Fig. 4.3. Curves of intraindividual and interindividual likehoods when four formants (in mel scale) were used as a features and GMM consist of 20 components

4.4 paveiksle b) pateikta gauta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 10 Gauso funkcijų sumos. Šiuo atveju gauta LKL taško reikšmė yra 12,59 %. Iš šių kreivių taip pat matome, kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 46 %. Taip pat matome, kai klaidingo „svetimo“ priėmimo tikimybė lygi 0 (KPL=0 %), tuomet KAL=97 %.

DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 15 Gauso funkcijų sumos, pateikta 4.4 paveiksle c). Gautoji LKL taško reikšmė yra 12,26 %. Iš šios kreivės taip pat matome, kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 50 %. Kai klaidingo „svetimo“ priėmimo tikimybė lygi 0 (KPL=0 %), tuomet KAL=96 %.



4.4 pav. DET kreivės, gautos požymiais naudojant keturias formantes melų skalėje ir panaudojant GMM komponenčių: a) 5; b) 10; c) 15; d) 20; e) adaptyvų skaičių

Fig. 4.4. DET curves when four formants in mel scale were used as features and count of GMM components: a) five; b) ten; c) fifteen; d) twenty; e) adaptive

4.4 paveiksle d) pateikta gauta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 20 Gauso funkcijų sumos. Gautoji LKL taško reikšmė yra 12,4 %. Taip pat matome, kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 51 %. Kai klaidingo „svetimo“ priėmimo tikimybė lygi 0 (KPL=0 %), tuomet KAL=96 %.

DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš adaptyvaus skaičiaus Gauso komponenčių, pateikta 4.4 paveiksle e). Šiuo atveju LKL=12,13 %. Kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 55,5 %. Taip pat matome, kai klaidingo „svetimo“ priėmimo tikimybė lygi 0 (KPL=0 %), tuomet KAL=96 %.

Iš atliktų eksperimentų matome, kad panaudojant požymiais keturias formantes (melų skalėje), gaunamas pakankamai didelis klaidų lygis, virš 12 %, nors ir dvigubai mažesnis nei požymiais panaudojant žadinimo signalo pagrindinį dažnį. Didinant mišinius sudarančių Gauso funkcijų skaičių atpažinimo tikslumas kinta nežymiai. Geriausias atpažinimo tikslumas gautas, kai panaudoti Gauso mišiniai, sudaryti iš adaptyvaus skaičiaus Gauso funkcijų sumos.

Atpažinimo rezultatai taip pat pateikti 4.2 lentelėje. Čia nurodytos LKL vertės, taip pat tikimybės, kad „savas“ bus atmestas kaip „svetimas“, kuomet „svetimo“ priėmimo tikimybė lygi 0, (KPL0) bei tikimybės, kad „svetimas“ bus priimtas kaip „savas“, kuomet „savojo“ atmetimo tikimybė lygi 0 (KAL0), esant skirtingam GMM komponenčių skaičiui.

Iš atliktų eksperimentų galima daryti išvadą, kad keturios formantės kaip požymiai gali būti panaudotos tose atpažinimo sistemose, kur nereikalingas didelis atpažinimo tikslumas, tačiau reikalingas didesnis skaičiavimo greitis bei griežtesni naudojamos atminties reikalavimai, kadangi šie požymių vektoriai susideda tik iš keturių komponenčių.

4.2 lentelė. Atpažinimo rezultatai panaudojus požymių sistemą, sudarytą iš keturių formančių

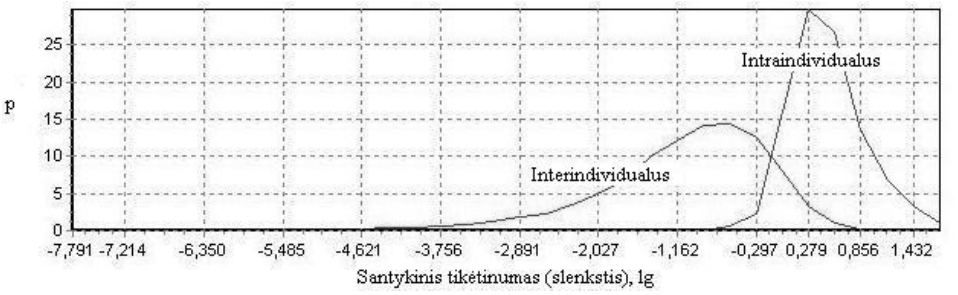
Table 4.2. Recognition results when four formants were used as features

GMM komponenčių skaičius	LKL, %	KPL0, % (KAL vertė kai KPL=0)	KAL0, % (KPL vertė kai KAL=0)
5	13,8	98	54,8
10	12,59	97	46
15	12,26	96	50
20	12,4	96	51
Adaptyvus	12,13	96	55,5

4.2.2.2. Kalbančiojo atpažinimas panaudojant keturias formantes bei tris antiformantes melų skalėje

Sekantis tyrimų etapas, formančių panaudojimas kartu su antiformantėmis. Tyrimų tikslas yra išsiaiškinti antiformančių įtaką atpažinimo tikslumui.

Mokymo metu gautos intraindividualių ir interindividualių tikėtinumų pasiskirstymų kreivės, kai panaudoti Gauso mišiniai, sudaryti iš 20 Gauso funkcijų (komponenčių) sumos, pavaizduotos 4.5 paveiksle.



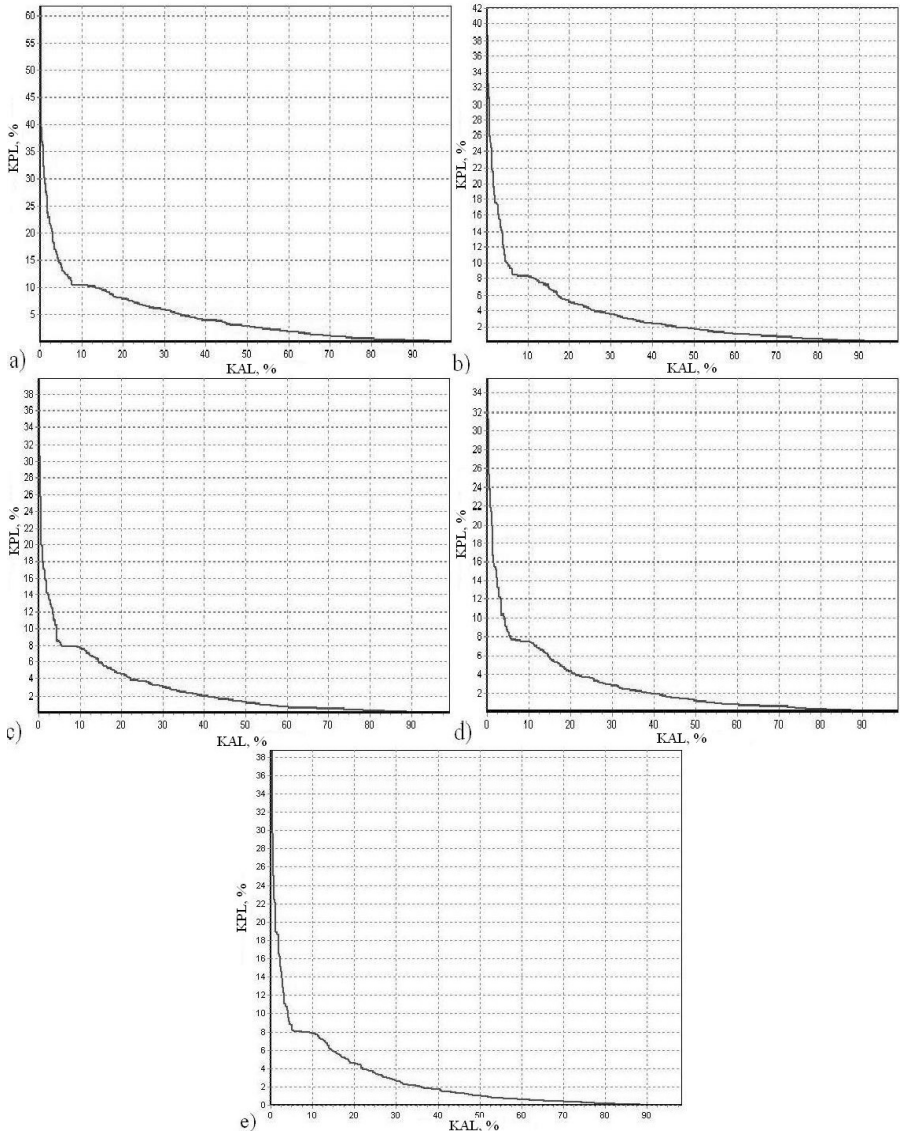
4.5 pav. Interindividualių ir intraindividualių tikėtinumų pasiskirstymų kreivės kai panaudoti požymiai yra keturios formantės bei trys antiformantės melų skalėje ir naudojant Gauso mišinius, sudarytus iš 20 Gauso funkcijų sumos

Fig. 4.5. Curves of intraindividual and interindividual likehoods when four formants and three antiformants in mel scale were used as features and GMM consist of 20 components

4.6 paveiksle pateiktos klaidingo „svetimo“ priėmimo tikimybės (KPL) priklausomybės nuo klaidingo „savojo“ atmetimo tikimybės (KAL) kreivė – DET, kai panaudoti Gauso mišiniai, sudaryti iš 5, 10, 15, 20 ir adaptyvaus skaičiaus Gauso funkcijų (komponenčių) sumos ir požymiais naudojant keturias formantes kartu su trimis antiformantėmis melų skalėje.

4.6 paveiksle a) pateikta gauta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 5 Gauso funkcijų sumos. Šiuo atveju gauta LKL taško reikšmė yra 10,35 %. Iš šios kreivės taip pat matome, kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 61,5 %. Kai klaidingo „svetimo“ priėmimo tikimybė lygi 0 (KPL=0 %), tuomet KAL=95 %.

4.6 paveiksle b) pateikta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 10 Gauso funkcijų sumos. Šiuo atveju gautoji LKL=8,33 %. Iš šių kreivių taip pat matome, kai KAL=0 %, tuomet KPL=42 %. Taip pat matome, kai KPL=0 %, tuomet KAL=92 %.



4.6 pav. DET kreivės, gautos požymiais naudojant keturias formantes bei tris antiformantes melų skalėje ir panaudojant GMM komponenčių: a) 5; b) 10; c) 15; d) 20; e) adaptyvų skaičių

Fig. 4.6. DET curves when four formants and three antiformants in mel scale were used as features and count of GMM components: a) five; b) ten; c) fifteen; d) twenty; e) adaptive

4.6 paveiksle c) pateikta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 15 Gauso funkcijų sumos. Šiuo atveju gautoji LKL reikšmė yra 7,99 %. Iš šių kreivių taip pat matome, kai KAL=0 %, KPL=40 %. Taip pat matome, kai KPL=0 %, tuomet KAL=88 %.

4.6 paveiksle d) pateikta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 20 Gauso komponenčių. Šiuo atveju gautoji LKL=7,62 %. Iš šių kreivių taip pat matome, kai KAL=0 %, KPL=35,5 %. Taip pat matome, kai KPL=0 %, tuomet KAL=88 %.

Atpažinimo rezultatai taip pat pateikti 4.3 lentelėje.

4.3 lentelė. Atpažinimo rezultatai panaudojus požymių sistemą, sudarytą iš keturių formančių bei trijų antiformančių melų skalėje

Table 4.3. Recognition results when four formants and three antiformants in mel scale were used as features

GMM komponenčių skaičius	LKL, %	KPL0, % (KAL vertė kai KPL=0)	KAL0, % (KPL vertė kai KAL=0)
5	10,35	95	61,5
10	8,33	92	42
15	7,99	88	40
20	7,62	88	35,5
Adaptyvus	8,01	90	38,5

LKL taško reikšmė, kai panaudoti Gauso mišiniai, sudaryti iš adaptyvaus skaičiaus Gauso komponenčių yra 8,01 % (4.6 paveikslas e)). Šiuo atveju, kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 38,5 %. Taip pat matome, kai klaidingo „svetimo“ priėmimo tikimybė lygi 0 (KPL=0 %), tuomet KAL=90 %.

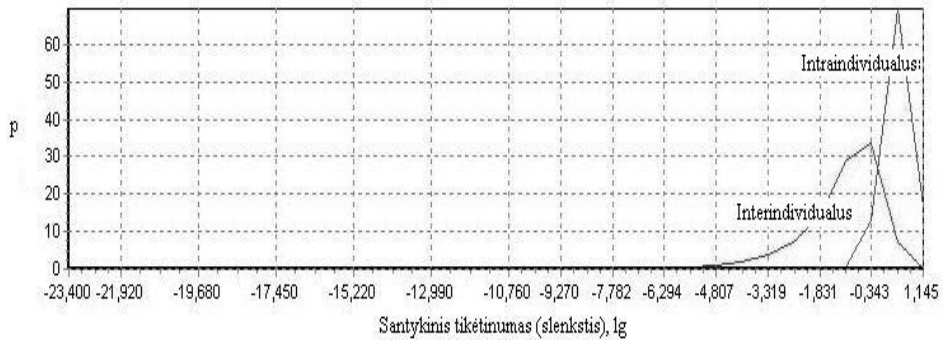
Iš atliktų eksperimentų matome, kad dėl papildomai panaudotų trijų antiformančių, atpažinimo rezultatai pagerėja virš 3%, lyginant su atpažinimo rezultatais, panaudojant tik keturias formantes. Reiktų paminėti, kad antiformantės nėra populiarios ir beveik nenaudojamos asmens atpažinimo sistemose, tačiau, kaip matome, jos turi nemažą įtaką atpažinimo tikslumui. Didėjant Gauso mišinių komponenčių skaičiui, atpažinimo tikslumas didėja.

4.2.3. Kalbančiojo atpažinimas panaudojant žadinimo signalo ir balso trakto parametrus

Toliau buvo atlikti kalbančiojo atpažinimo tyrimai, kartu derinant balso trakto ir žadinimo signalo požymius. Požymiais buvo naudojama: keturios formantės kartu su žadinimo signalo pagrindiniu dažniu bei keturios formantės, trys antiformantės ir žadinimo signalo pagrindinis dažnis. Formantės bei antiformantės buvo skaičiuojamos melų skalėje. Eksperimentai atlikti panaudojant Gauso mišinius, sudarytus iš 5, 10, 15, 20 bei adaptyvaus Gauso funkcijų, skaičiaus.

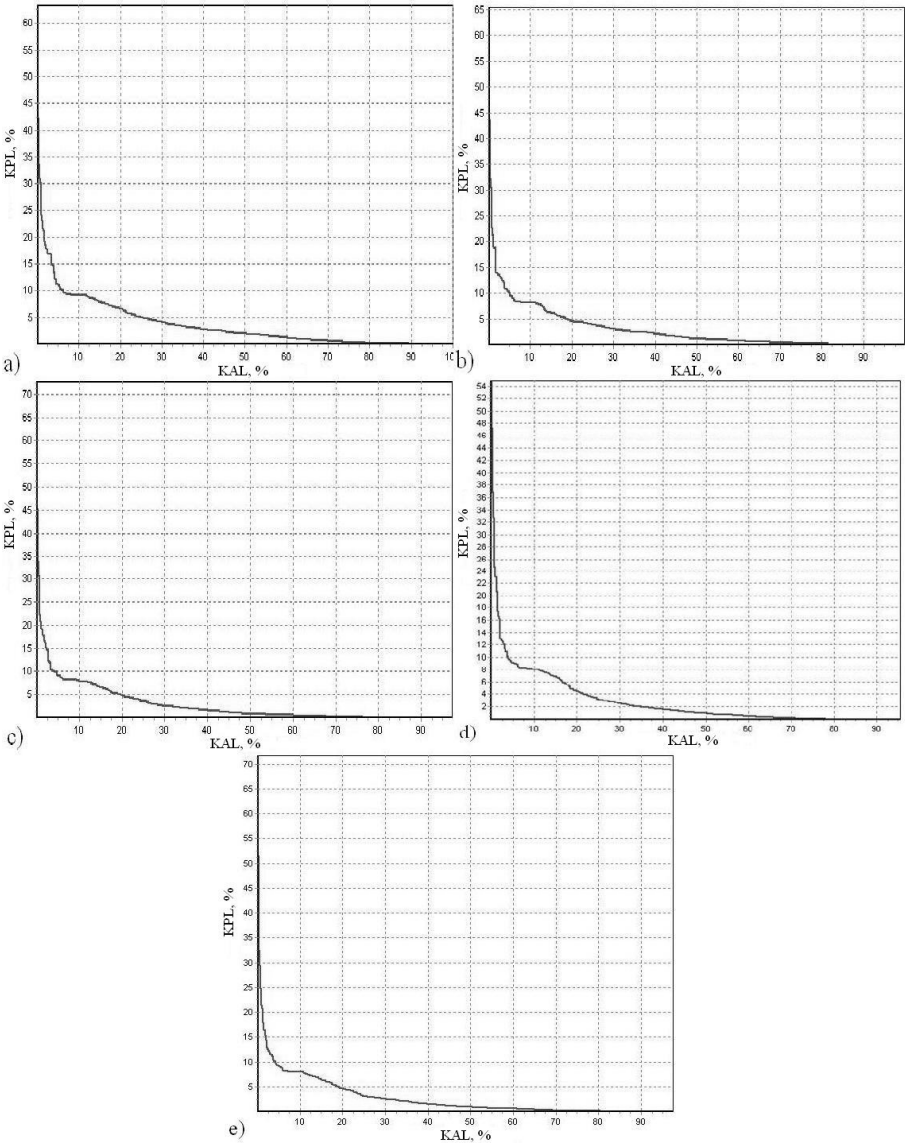
4.2.3.1. Kalbančiojo atpažinimas panaudojant keturias formantes su žadinimo signalo pagrindiniu dažniu

Mokymo metu gautos intraindividualių ir interindividualių tikėtinumų pasiskirstymų kreivės, kai panaudoti Gauso mišiniai, sudaryti iš 20 Gauso funkcijų sumos, pavaizduotos 4.7 paveiksle.



4.7 pav. Interindividualių ir intraindividualių tikėtinumų pasiskirstymų kreivės kai panaudoti požymiai yra keturios formantės melų skalėje su F0 ir parenkant Gauso mišinius, sudarytus iš 20 Gauso funkcijų sumos

Fig. 4.7. Curves of intraindividual and interindividual likehoods when pitch and four formants in mel scale were used as features and GMM consist of 20 components



4.8 pav. DET kreivės, gautos požymiais naudojant keturias formantes melų skalėje kartu su F0 ir panaudojant GMM komponenčių: a) 5; b) 10; c) 15; d) 20; e) adaptyvų skaičių

Fig. 4.8. DET curves when pitch and four formants in mel scale were used as features and count of GMM components: a) five; b) ten; c) fifteen; d) twenty; e) adaptive

DET kreivės, kai panaudoti Gauso mišiniai, sudaryti iš 5, 10, 15, 20 ir adaptyvaus skaičiaus Gauso mišinių komponenčių ir požymiais naudojant keturias formantes kartu su žadinimo signalo pagrindiniu dažniu (F0), pateiktos 4.8 paveiksle.

DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 5 Gauso funkcijų sumos, pateikta 4.8 paveiksle a). Kaip matome, šiuo atveju gauta LKL=9,22 %. Kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 63,5 %, kai klaidingo „svetimo“ priėmimo tikimybė lygi 0 (KPL=0 %), tuomet KAL=90 %.

4.8 paveiksle b) pateikta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 10 Gauso funkcijų sumos. Šiuo atveju gautoji LKL reikšmė yra 8,25 %. Iš šios kreivės taip pat matome, kai KAL=0 %, tuomet KPL=65,8 %. Taip pat matome, kai KPL=0 %, tuomet KAL=85 %.

4.8 paveiksle c) pateikta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 15 Gauso funkcijų sumos. Gautoji LKL reikšmė yra 8,2 %. Iš šios kreivės taip pat matome, kai KAL=0 %, tuomet KPL=72,5 %. Taip pat matome, kai klaidingo „svetimo“ priėmimo tikimybė lygi 0 (KPL=0 %), tuomet KAL=85 %.

4.8 paveiksle d) pateikta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 20 Gauso komponenčių sumos. Gautoji LKL taško reikšmė yra 8,17 %. Šiuo atveju, kai KAL=0 %, tuomet KPL=55 %. Kai klaidingo „svetimo“ priėmimo tikimybė artima 0 (KPL≈ 0 %), tuomet KAL≈ 80 %.

DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš adaptyvaus Gauso komponenčių skaičiaus. pateikta 4.8 paveiksle e). Gautoji LKL taško reikšmė yra 8,13 %. Iš šių kreivių taip pat matome, kai KAL=0 %, tuomet KPL=72 %. Taip pat matome, kai klaidingo „svetimo“ priėmimo tikimybė artima 0 (KPL≈ 0 %), tuomet KAL≈ 80 %. Atpažinimo rezultatai pateikti ir 4.4 lentelėje.

4.4 lentelė. Atpažinimo rezultatai panaudojus požymių sistemą, sudarytą iš keturių formančių melų skalėje bei F0

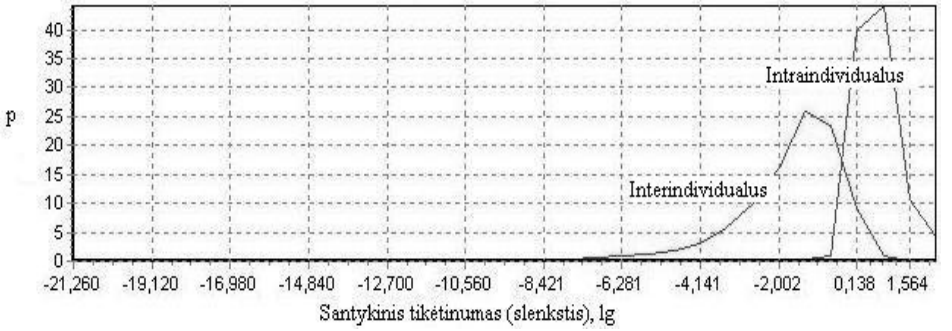
Table 4.4. Recognition results when pitch and four formants in mel scale were used

GMM komponenčių skaičius	LKL, %	KPL0, % (KAL vertė kai KPL=0)	KAL0, % (KPL vertė kai KAL=0)
5	9,22	90	63,5
10	8,25	85	65,8
15	8,2	85	72,5
20	8,17	80	55
Adaptyvus	8,13	80	72

Kaip matome iš šių eksperimentų, apjungus žadinimo signalo ir balso trakto požymius gaunami geresni rezultatai. Taip pat reiktų pastebėti, kad panaudojant keturias formantes kartu su trimis antiformantėmis gaunami geresni atpažinimo rezultatai nei keturias formantes kartu su F0, nors šis skirtumas nėra didelis.

4.2.3.2. Kalbančiojo atpažinimas panaudojant keturias formantes, tris antiformantes ir žadinimo signalo pagrindinį dažnį

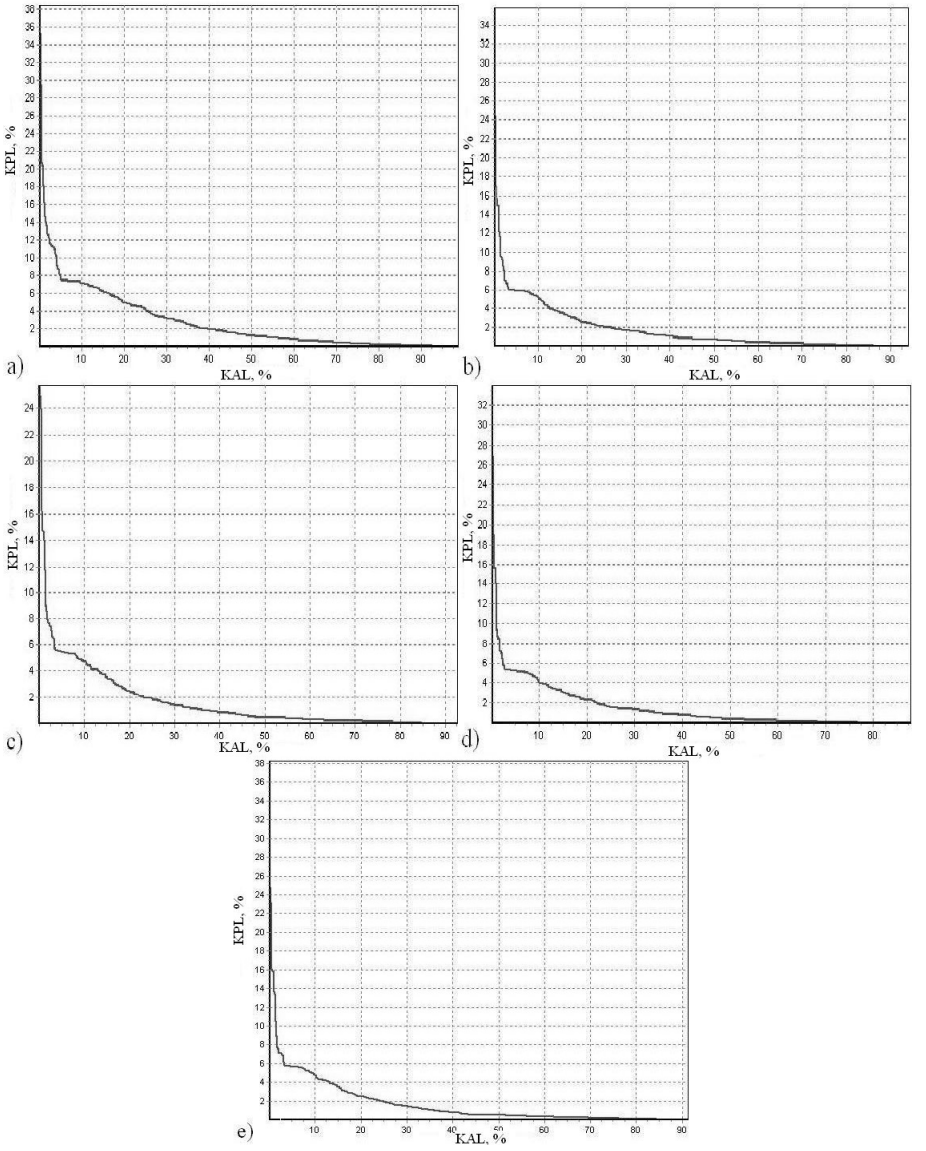
Tolesnis tyrimų etapas, panaudoti pasiūlytą požymių sistemą: keturias formantes, tris antiformantes, melų skalėje, bei F0. Mokymo metu gautos intraindividualių ir interindividualių tikėtinumų pasiskirstymų kreivės, kai panaudoti Gauso mišiniai, sudaryti iš 20 Gauso funkcijų sumos, pavaizduotos 4.9 paveiksle.



4.9 pav. Interindividualių ir intraindividualių tikėtinumų pasiskirstymų kreivės kai panaudoti požymiai yra keturios formantės trys antiformantės bei žadinimo signalo pagrindinis dažnis ir panaudojant 20 Gauso mišinių komponenčių

Fig. 4.9. Curves of intraindividual and interindividual likehoods when pitch four formants and three antiformants in mel scale were used as a features and GMM consist of 20 components

DET kreivės, kai panaudoti Gauso mišiniai, sudaryti iš 5, 10, 15, 20 ir adaptyvaus skaičiaus Gauso mišinių komponenčių ir požymiais naudojant keturias formantes, tris antiformantes kartu su žadinimo signalo pagrindiniu dažniu (F0), pateiktos 4.10 paveiksle.



4.10 pav. DET kreivės, gautos požymiais naudojant keturias formantes kartu su trimis antiformantėmis melų skalėje bei F0 ir panaudojant GMM komponenčių: a) 5; b) 10; c) 15; d) 20; e) adaptyvų skaičių

Fig. 4.10. DET curves when pitch four formants three antiformants in mel scale were used as features and count of GMM components: a) five; b) ten; c) fifteen; d) twenty; e) adaptive

4.5 lentelėje pateikti atpažinimo sistemos tikslumo parametrai, esant skirtingam GMM komponenčių skaičiui.

4.5 lentelė. Atpažinimo rezultatai panaudojus pasiūlytą požymių sistemą

Table 4.5. Recognition results when proposed system of features was used

GMM komponenčių skaičius	LKL, %	KPL0, % (KAL vertė kai KPL=0)	KAL0, % (KPL vertė kai KAL=0)
5	7,44	93	38,2
10	5,94	86	36
15	5,45	85	25,8
20	5,17	78	34
Adaptyvus	5,66	85	38,2

Iš atliktų eksperimentų matome, kad panaudojus pasiūlytą požymių sistemą, apjungus žadinimo šaltinio ir balso trakto požymius, gauti tiksliausi atpažinimo rezultatai. Kai panaudota 20 GMM komponenčių, LKL vertė gauta 5,17 %. Didėjant GMM komponenčių skaičiui, atpažinimo tikslumas didėja. Didžiausias atpažinimo tikslumas pasiektas panaudojus 20 GMM komponenčių.

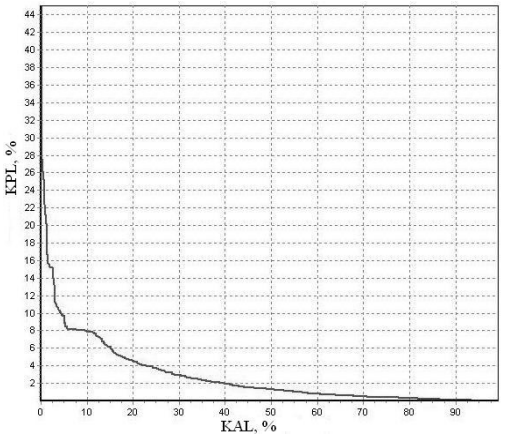
4.2.4. Kalbančiojo atpažinimo tyrimas skaičiuojant formantes bei antiformantes tiesinėje ir melų skalėje

Kadangi aukštesnių formančių dispersija didesnė nei žemesnių, pasiūlėme skaičiuoti formantes bei antiformantes melų skalėje, kuri iki 1 kHz yra beveik tiesinė, toliau logaritminė. Tokiu būdu „sulyginama“ aukštesnių ir žemesnių formančių bei antiformančių dispersija. Toliau bus pateikti eksperimentų rezultatai, požymiais naudojant keturias formantes kartu su trimis antiformantėmis bei keturias formantes kartu su žadinimo signalo pagrindiniu dažniu. Šiuose dviejuose eksperimentuose formantės ir antiformantės skaičiuojamos tiesinėje dažnių skalėje.

4.2.4.1. Kalbančiojo atpažinimo tyrimas panaudojant keturias formantes ir tris antiformantes tiesinėje skalėje

4.11 paveiksle pateikta gauta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 15 GMM komponenčių.

Gautoji LKL (lygių klaidų lygio) taško reikšmė yra 8,1 %.



4.11 pav. DET kreivė, gauta panaudojant 4 formantes bei 3 antiformantes tiesinėje skalėje ir naudojant 15 GMM komponenčių

Fig. 4.11. DET curve when 4 formants and 3 antiformants in linear scale were used as features and 15 GMM components were taken

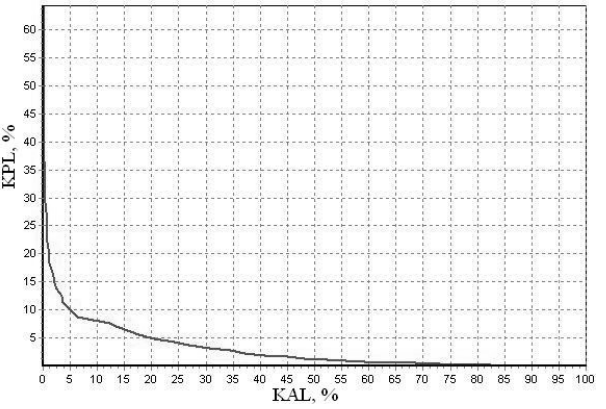
Iš šios kreivės taip pat matome, kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 45 % (t. y. KAL0=45 %). Taip pat matome, kai klaidingo „svetimo“ priėmimo tikimybė artima 0 (KPL≈ 0 %), tuomet KAL≈ 95 %.

Lyginant šią reikšmę su LKL reikšme, panaudojus keturias formantes su trimis antiformantėmis melų skalėje, kuri gavosi 7,99 % (4.6 paveikslas c)), matome, kad formančių ir antiformančių skaičiavimas melų skalėje šiek tiek pagerino atpažinimo rezultatus.

4.2.4.2. Kalbančiojo atpažinimo tyrimas panaudojant keturias formantes tiesinėje skalėje bei žadinimo signalo pagrindinį dažnį

4.12 paveiksle pateikta gauta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš 10 Gauso funkcijų sumos ir požymiais panaudotos keturios formantės tiesinėje skalėje bei žadinimo signalo pagrindinis dažnis. Gautoji LKL reikšmė yra 8,51 %.

Iš šios kreivės taip pat matome, kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 64 %. Taip pat matome, kai klaidingo „svetimo“ priėmimo tikimybė artima 0 (KPL≈ 0 %), tuomet KAL≈ 90 %.



4.12 pav. DET kreivė, gauta panaudojant 4 formantes tiesinėje skalėje kartu su žadinimo signalo pagrindiniu dažniu ir naudojant Gauso mišinius, sudarytus iš 10 Gauso funkcijų sumos.

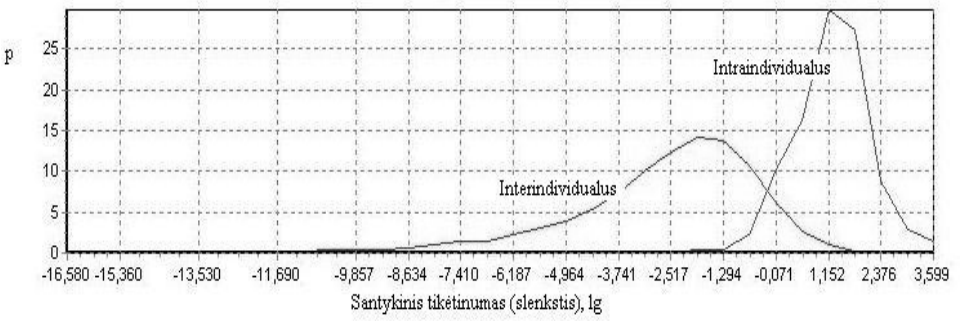
Fig. 4.12. DET curves when pitch and four formants in linear scale were used as features and 10 GMM components were used

Lyginant šią LKL reikšmę su LKL reikšme, gauta panaudojus keturias formantes melų skalėje su žadinimo signalo pagrindiniu dažniu, paėmus Gauso mišinius, sudarytus iš 10 Gauso funkcijų, kuri gavosi 8,25 % (4.8 paveikslas b)), matome, kad vėl formančių skaičiavimas melų skalėje šiek tiek pagerino atpažinimo rezultatus.

4.2.5. Kalbančiojo atpažinimo tyrimas standartinius melų skalės kepstro koeficientus

Kad palygintume pasiūlytos požymių sistemos efektyvumą, buvo atlikti kalbančiojo atpažinimo tyrimai panaudojant standartinius melų skalės kepstro koeficientus (MSKK). Skaičiuojant MSKK buvo panaudota 25 trikampiai filtrai melų skalės spektro (MSSK) skaičiavimui, bei iš jo, pritaikius DKT, apskaičiuota 13 pirmųjų MSKK koeficientų. Eksperimentai atlikti panaudojant Gauso mišinius, sudarytus iš 5, 10, 15, 20 bei adaptyvaus Gauso funkcijų, skaičiaus, sumos.

Mokymo metu gautos intraindividualių ir interindividualių tikėtinumų pasiskirstymų kreivės, kai panaudoti Gauso mišiniai, sudaryti iš 20 Gauso funkcijų sumos, pavaizduotos 4.13 paveiksle.



4.13 pav. Interindividualių ir intraindividualių tikėtinumų pasiskirstymų kreivės kai panaudoti požymiai yra 13 eilės standartiniai MSKK ir naudojant Gauso mišinius, sudarytus iš 20 komponenčių

Fig. 4.13. Curves of intraindividual and interindividual likehoods when standard MFCC were used as features and GMM consisted of 20 components

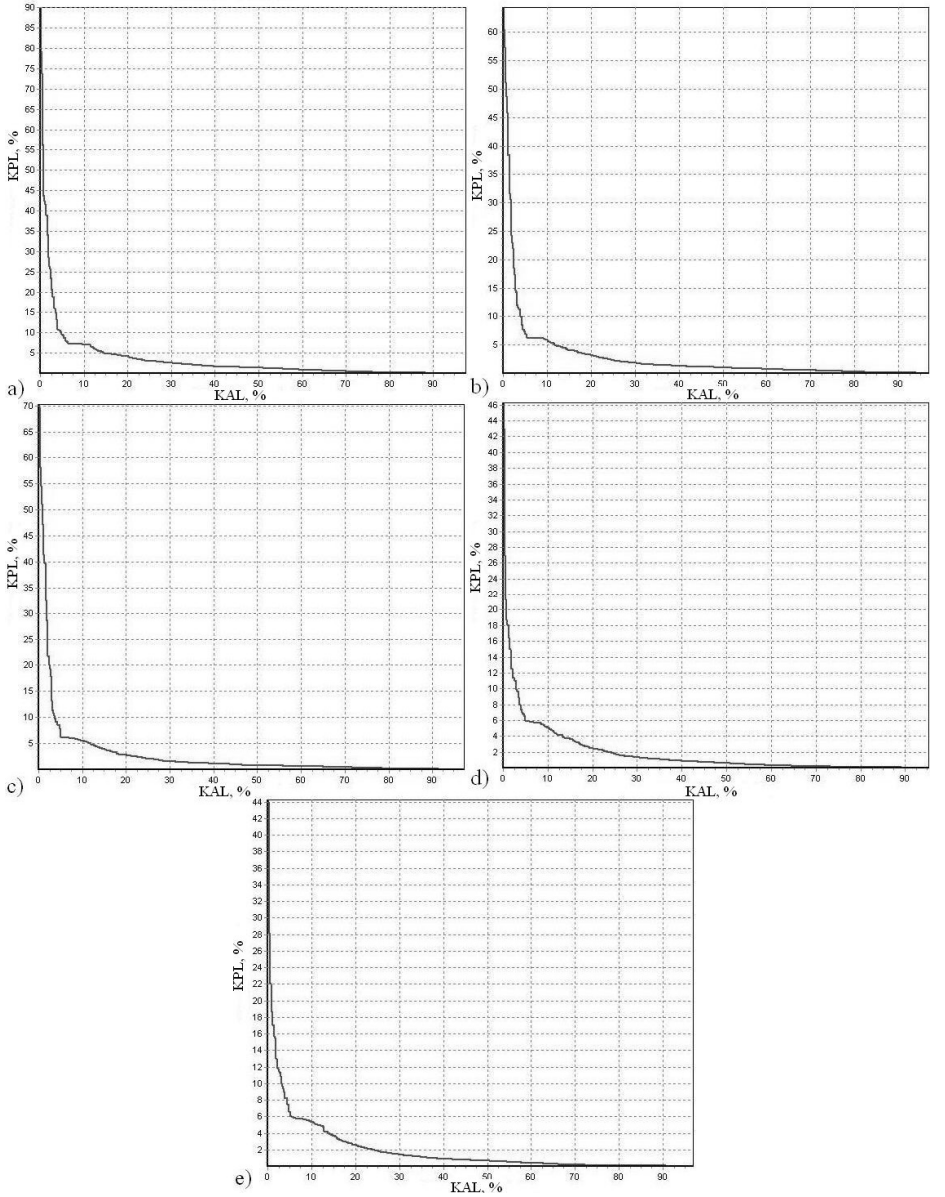
4.6 lentelėje pateikti atpažinimo sistemos tikslumo parametrai, esant skirtingam GMM komponenčių skaičiui. Lentelėje nurodoma LKL (lygių klaidų lygio) vertė, taip pat tikimybė, kad „savas“ bus atmestas kaip „svetimas“, (pažymėta KPL0), kuomet „savojo“ atmetimo tikimybė lygi 0, t. y. KPL=0 % bei tikimybė, kad „svetimas“ bus priimtas kaip „savas“, (pažymėta KAL0), kuomet „savojo“ atmetimo tikimybė lygi 0, t. y. KAL=0 %.

4.6 lentelė. Atpažinimo rezultatai panaudojus standartinius MSKK

Table 4.6. Recognition results when standard MFCC were used

GMM komponenčių skaičius	LKL, %	KPL0, % (KAL vertė kai KPL=0)	KAL0, % (KPL vertė kai KAL=0)
5	7,35	97	90
10	6,27	97	64,5
15	6,15	95	70
20	5,86	95	46
Adaptyvus	5,89	95	44

DET kreivės, kai panaudoti Gauso mišiniai, sudaryti iš 5, 10, 15, 20 ir adaptyvaus skaičiaus Gauso mišinių komponenčių ir požymiais naudojant standartinius MSKK, pateiktos 4.14 paveiksle.



4.14 pav. DET kreivės, gautos požymiais imant standartinius MSKK ir panaudojant GMM komponenčių: a) 5; b) 10; c) 15; d) 20; e) adaptyvų skaičių

Fig. 4.14. DET curves when standard MFCC were used as features and count of GMM components: a) five; b) ten; c) fifteen; d) twenty; e) adaptive

4.2.6. Atpažinimo tikslumo priklausomybė nuo pradinių GMM parametrų vertinimo algoritmo

Toliau buvo atlikti kalbančiojo atpažinimo tyrimai panaudojant skirtingus pradinių GMM parametrų vertinimo algoritmus. Klasteriai, skirti GMM pradinių parametrų vertinimui, buvo sudaromi trimis būdais:

1. Panaudojant modifikuotą vektorinio kvantavimo (VK) algoritmą.
2. Tiesiškai dalinant požymių vektorius į klasterius.
3. Atsitiktinis klasterių formavimas.

Požymiais buvo parinkta pasiūlyta požymių vektorių sistema, susidedanti iš 4 formančių ir 3 antiformančių melų skalėje ir žadinimo signalo pagrindinio dažnio. Gauso funkcijų, sudarančių Gauso mišinius, skaičius buvo parenkamas adaptyviai, priklausomai nuo modelio kūrimui skirtų požymių vektorių skaičiaus ir buvo lygus požymių vektorių skaičiui padalintam iš 100, apvalinant į didesnę pusę. Eksperimentų tikslas – nustatyti ir palyginti kalbėtojo GMM kūrimui reikalingų iteracijų skaičių ir atpažinimo tikslumo priklausomybę nuo pradinių parametrų vertinimo algoritmo. 4.7 lentelėje pateikti iteracijų skaičiai, panaudoti tikslinant tirtų kalbėtojų modelių parametrus, taip pat ir tų kalbėtojų modelių kūrimui panaudotų požymių vektorių skaičiai bei GMM komponenčių skaičiai. Nurodomos ir gautos LKL vertės.

4.7 lentelė. Iteracijų skaičiai skirtingai vertinant pradinius GMM parametrus

Table 4.7. Count of iterations when different methods for estimation of initial GMM parameters were used

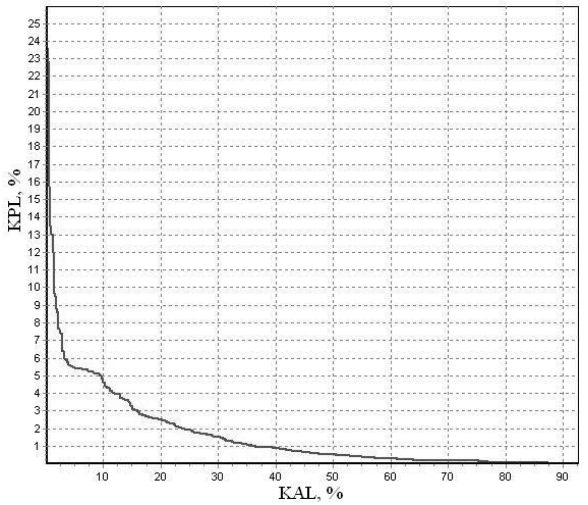
Kalbėtojas	Pradinio GMM parametrų vertinimo algoritmas ir iteracijų skaičius			Vektorių skaičius	GMM komponenčių skaičius
	VK	Tiesinis	Atsitiktinis		
1	2	3	4	5	6
01	53	50	27	1474	15
02	30	28	62	1495	15
03	48	43	27	1383	14
05	42	33	35	1452	15
06	46	36	40	1378	14
07	14	43	29	979	10
08	41	26	46	1375	14
11	49	70	51	1424	15
12	25	52	39	1519	16

4.7 lentelės pabaiga

1	2	3	4	5	6
13	23	41	44	1274	13
14	47	32	45	1305	14
15	21	20	29	926	10
16	36	29	43	1408	15
17	54	30	29	1303	14
18	28	31	24	1163	12
19	30	43	19	1281	13
21	20	23	22	1109	12
24	20	30	43	1245	13
27	55	28	27	1338	14
28	57	40	30	1306	14
29	66	46	49	1636	17
31	58	43	33	1112	12
32	35	48	30	981	10
33	39	25	40	1592	16
34	43	40	32	1163	12
36	34	24	36	1193	12
37	23	28	24	1034	11
38	59	45	43	1312	14
50	36	27	53	1107	12
52	34	27	53	1473	15
53	52	26	61	1337	14
64	22	34	27	998	10
66	49	35	29	1268	13
74	36	32	52	1346	14
75	22	24	55	1572	10
76	60	24	38	1016	11
78	40	37	26	1025	11
81	33	44	44	1636	17
82	37	37	41	1353	14
84	31	36	33	1061	11
88	14	39	34	662	7
LKL, %	5,39	6,27	6,1		

Pažvelgę į 4.7 lentelę mes matome, kad nei vienas iš panaudotų pradinio GMM parametrų vertinimo algoritmų nedavė išskirtinai mažo ar didelio reikalingų iteracijų skaičiaus, visų jų skaičius svyruoja gana įvairiai. Bet kaip matyti, nuo to priklauso atpažinimo tikslumo rezultatai. Vertinant pradinius parametrus modifikuotu LBG vektorinio kvantavimo algoritmu buvo gauti geriausi atpažinimo rezultatai, t. y. mažiausia LKL vertė.

4.15 paveiksle pateikta šių eksperimentų metu gauta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš adaptyvaus komponenčių skaičiaus ir pradiniam parametrų vertinimui panaudotas modifikuotas VK algoritmas. Gautoji LKL (lygių klaidų lygio) taško reikšmė yra 5,39 %. Iš šių kreivių taip pat matome, kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 26 %. Taip pat matome, kai klaidingo „svetimo“ priėmimo tikimybė artima 0 (KPL≈ 0 %), tuomet KAL≈ 80 %.

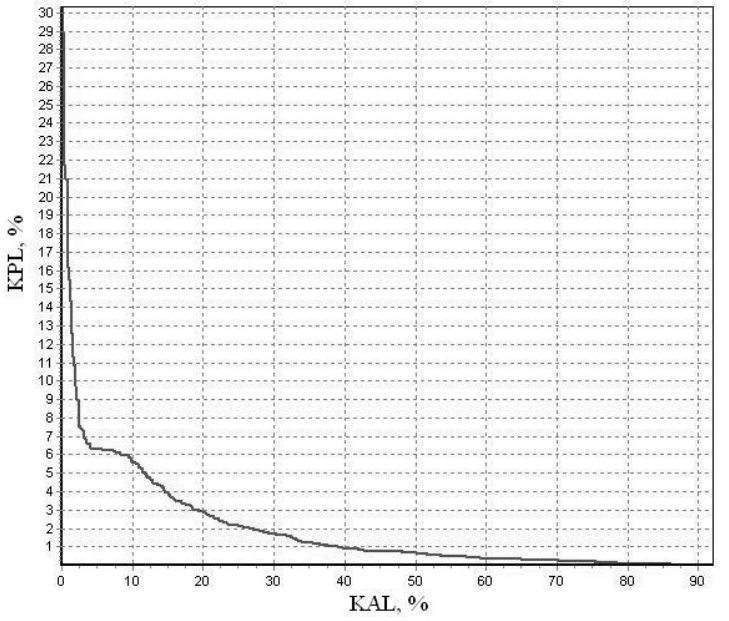


4.15 pav. DET kreivė, gauta panaudojant pasiūlytą požymių sistemą bei Gauso mišinius, sudarytus iš adaptyvaus komponenčių skaičiaus. Pradinis parametrų vertinimas atliktas panaudojant modifikuotą VK algoritmą

Fig. 4.15. DET curves when proposed system of features was used and count of GMM components was adaptive. Modified VQ method was used for GMM parameter initialization

4.16 paveiksle pateikta šių eksperimentų metu gauta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš adaptyvaus komponenčių skaičiaus ir

pradiniam parametrų vertinimui panaudotas tiesinio dalijimo į klasterius algoritmas.



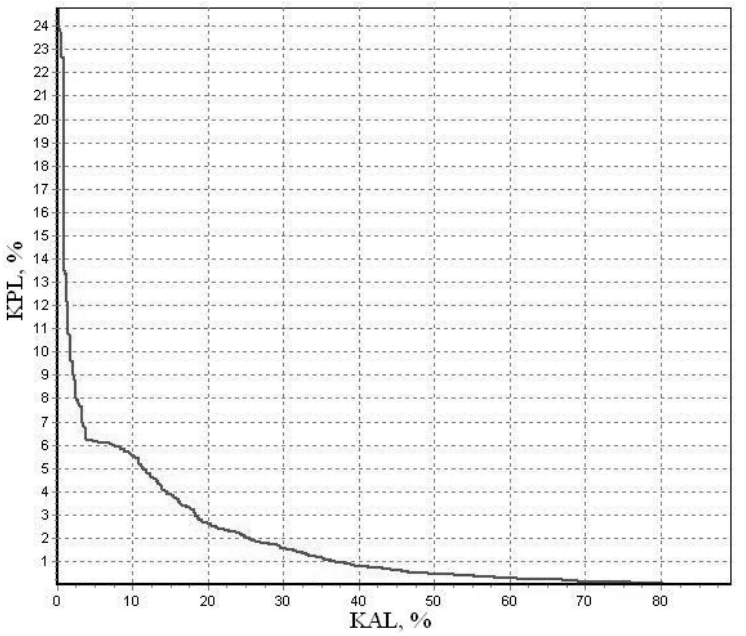
4.16 pav. DET kreivė, gauta panaudojant pasiūlytą požymių sistemą bei Gauso mišinius, sudarytus iš adaptyvaus komponenčių skaičiaus. Pradinis parametrų vertinimas atliktas panaudojant tiesinio dalijimo į klasterius algoritmą

Fig. 4.16. DET curves when proposed system of features was used and count of GMM components was adaptive. Method of linear splitting of clusters was used for GMM parameter initialization

Gautoji LKL (lygių klaidų lygio) taško reikšmė yra 6,27 %.

Iš šios kreivės taip pat matome, kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 30 %. Taip pat matome, kai klaidingo „svetimo“ priėmimo tikimybė artima 0 (KPL≈ 0 %), tuomet KAL≈ 90 %.

4.17 paveiksle pateikta šių eksperimentų metu gauta DET kreivė, kai panaudoti Gauso mišiniai, sudaryti iš adaptyvaus komponenčių skaičiaus ir pradiniam parametrų vertinimui panaudotas atsitiktinio klasterių formavimo algoritmas.



4.17 pav. DET kreivė, gauta panaudojant pasiūlytą požymių sistemą bei Gauso mišinius, sudarytus iš adaptyvaus komponenčių skaičiaus. Pradinis parametrų vertinimas atliktas panaudojant atsitiktinio klasterių formavimo algoritmą

Fig. 4.17. DET curves when proposed system of features was used and count of GMM components was adaptive. Method of random forming of clusters was used for GMM parameter initialization

Gautoji LKL taško reikšmė yra 6,1 %.

Iš šių kreivių taip pat matome, kai klaidingo „savojo“ atmetimo (KAL) tikimybė lygi 0, klaidingo „svetimo“ priėmimo (KPL) tikimybė lygi 25 %. Taip pat matome, kai klaidingo „svetimo“ priėmimo tikimybė artima 0 (KPL≈ 0 %), tuomet KAL≈ 90 %.

4.3. Atpažinimo rezultatų apibendrinimas

Atliktų atpažinimo eksperimentų pagrindiniai rezultatai pateikti 4.8 lentelėje, kur nurodomos gautos LKL vertės panaudojus skirtingas požymių sistemas bei skirtingus GMM komponenčių skaičių.

4.8 lentelė. Atliktų kalbančiojo atpažinimo eksperimentų rezultatai

Table 4.8. Experimental results of speaker recognition

GMM komponenčių skaičius	Požymiai					
	F0	4F	4F3A	4FF0	**4F3AF0**	**MSKK**
1	26,65	-	-	-	-	-
3	24,62	-	-	-	-	-
5	24,71	13,8	10,35	9,22	7,44	7,35
10	24,6	12,59	8,33	8,25	5,94	6,27
15	-	12,26	7,99	8,2	**5.45**	6,15
20	-	12,4	7,62	8,17	**5,17**	5,86
Adaptyvus	-	12,13	8,01	8,13	**5,66**	5,89

Čia F0 žymėjimas atitinka žadinimo signalo pagrindinį dažnį, 4F žymi požymių sistemą, susidedančią iš keturių formančių, 4F3A žymi požymių sistemą, susidedančią iš keturių formančių ir trijų antiformančių, 4FF0 žymi požymių sistemą, susidedančią iš keturių formančių ir žadinimo signalo pagrindinio dažnio, 4F3AF0 žymi požymių sistemą, susidedančią iš keturių formančių, trijų antiformančių ir žadinimo signalo pagrindinio dažnio.

Kaip matome iš šios lentelės, geriausi atpažinimo rezultatai (mažiausia LKL vertė) buvo gauti panaudojus pasiūlytą požymių vektorių sistemą, susidedančią iš 4 formančių, 3 antiformančių ir žadinimo signalo pagrindinio dažnio.

Taip pat, žvelgiant į pateiktas DET kreives, galima vertinti ir kitas atpažinimo sistemos tikslumo charakteristikas. Mažiausia klaidingo „svetimo“ priėmimo tikimybė (KPL) kuomet klaidinga „savojo“ atmetimo tikimybė lygi 0, (KAL=0) taip pat buvo gauta panaudojus pasiūlytą požymių vektorių sistemą, susidedančią iš 4 formančių, 3 antiformančių ir žadinimo signalo pagrindinio dažnio, kuri buvo lygi apie 25 %.

Mažiausia klaidingo „savojo“ atmetimo tikimybė (KAL), kuomet klaidinga „svetimo“ priėmimo tikimybė lygi 0 (KPL=0), buvo gauta taip pat panaudojus pasiūlytą požymių vektorių sistemą, susidedančią iš 4 formančių, 3 antiformančių ir žadinimo signalo pagrindinio dažnio, kuri buvo lygi apie 87 %.

4.4. Ketvirtojo skyriaus apibendrinimas

- Naudojant sukurtą kalbančiojo atpažinimo sistemą, atlikti kalbančiojo atpažinimo tyrimai, panaudojant įvairias požymių vektorių sistemas: žadinimo signalo dažnį; keturias formantes; keturias formantes kartu su

trimis antiformantėmis; keturias formantes su žadinimo signalo dažniu; keturias formantes kartu su trimis antiformantėmis ir žadinimo signalo dažniu; standartinius MSKK.

- Ištirta atpažinimo tikslumo priklausomybė nuo Gauso mišinius sudarančių Gauso funkcijų skaičiaus, panaudojant įvairias požymių vektorių sistemas.

- Atlikti kalbančiojo atpažinimo tyrimai, panaudojant skirtingus pradinių GMM parametrų vertinimo algoritmus.

- Tirti formančių bei antiformančių skaičiavimo tiesinėje ir melų skalėje atvejai.

- Geriausi atpažinimo rezultai buvo gauti panaudojus pasiūlytą požymių sistemą, susidedančią iš keturių formančių, trijų antiformančių ir žadinimo signalo pagrindinio dažnio.

- Atpažinimo tikslumas taip pat priklauso ir nuo pradinio GMM parametrų vertinimo. Tiksliausi atpažinimo rezultatai pasiekti tam tikslui panaudojant modifikuotą vektorinio kvantavimo algoritmą.

Bendrosios išvados

Darbo metu buvo atlikta kalbančiojo atpažinimo sistemų analizė, atliktas kalbančiojo atpažinimo, panaudojant pasiūlytus sprendimus, tyrimas. Gautus darbo rezultatus apibendriname ir pateikiame išvadas:

1. Pasiūlytas automatinis vokalizuotų garsų išrinkimo (segmentavimo) metodas, nereikalaujantis iš vartotojo jokių triukšmo ir kalbos signalo pavyzdžių nurodymo. Foninio triukšmo parametrai automatiškai nustatomi atmetus visus kadrus su nulinėmis ir labai žemomis signalo reikšmėmis ir po to randant tam tikrą kadrų skaičių su minimaliomis melų skalės spektro energijos reikšmėmis.

2. Pasiūlyta požymių sistema, skirta asmens atpažinimui pagal balsą, susidedanti iš žadinimo signalo parametrų bei balso trakto parametrų. Kaip žadinimo signalo parametrą panaudojome balso stygų virpėjimo dažnį – žadinimo signalo pagrindinį dažnį F0. Kaip balso trakto parametrus panaudojome keturias formantes (kalbos signalo spektro gaubtinės maksimumų dažnius) bei tris antiformantes (kalbos signalo spektro gaubtinės minimumų dažnius). Gauti atpažinimo rezultatai pagal visus tikslumo parametrus pralenkė šiuo metu vienus iš plačiausiai naudojamų pasaulyje spektrinių požymių – melų skalės kepstro koeficientus (MSKK), naudojamus kalbos bei asmens atpažinime. Gautas lygių klaidų lygis panaudojant pasiūlytą požymių vektorių sistemą – LKL=5,17 %, tuo tarpu, panaudojus MSKK, LKL=5,86 %.

3. Pasiūlytos požymių sistemos vektoriai turi mažesnį komponenčių skaičių, vektorius susideda iš 8 komponenčių, tuo tarpu, panaudojant 13 eilės MSKK,

vektorius susideda iš 13 komponenčių. Dėl šių priežasčių tiek kuriant kalbėtojų modelius, tiek ir atpažinimo metu, naudojant pasiūlytą požymių vektorių sistemą reikia atlikti apie 1,6 karto mažiau skaičiavimo operacijų. Dėl to teigiame, kad požymių sistema, susidedanti iš 4 formančių, 3 antiformančių bei žadinimo signalo pagrindinio dažnio F0 gali būti panaudota kalbančiojo atpažinimui pagal balsą ir tai yra efektyvesnė požymių sistema, nei standartinė – MSKK.

4. Kadangi aukštesnių formančių bei antiformančių dispersija yra didesnė nei žemesnių, kad „suvienodinti“ dispersijas, pasiūlėme formantes bei antiformantes skaičiuoti melų skalėje. Atlikus eksperimentus, paaiškėjo, kad formančių bei antiformančių skaičiavimas melų skalėje šiek tiek pagerino (LKL reikšmė sumažėjo 0,11–0,26 %) atpažinimo tikslumą, nei naudojant jas tiesinėje skalėje. Todėl teigiame, kad formantes bei antiformantes geriau skaičiuoti melų skalėje.

5. Pasiūlytas metodas pradinių GMM parametrų vertinimui. Tam tikslui panaudotas modifikuotas LBG vektorinio kvantavimo (VK) algoritmas. Atlikus eksperimentus paaiškėjo, kad panaudojus pradinių parametrų vertinimui VK metodą, gautas didžiausias atpažinimo tikslumas (lygių klaidų lygis sumažėjo 0,71–0,88 %), nei panaudojant atsitiktinio klasterių formavimo ar tiesiško požymių vektorių dalijimo į klasterius metodus. Tačiau VK metodas nesumažino iteracijų skaičiaus, reikalingo tikslinant GMM parametrus. Todėl galima teigti, kad vertinant pradinius GMM parametrus geriau naudoti vektorinio kvantavimo metodą, nei parinkti atsitiktines parametrų vertes.

Disertacijos darbo rezultatai parodė tolimesnę atpažinimo sistemos vystymo kryptį – esamos požymių sistemos gerinimą bei jos papildymą panaudojant kitus spektrinius bei žadinimo šaltinio požymius. Papildomais kalbančiojo požymiais gali būti panaudotas balso tembras, pirmųjų formančių amplitudžių santykis ir t. t.

Literatūros sąrašas

Andrews, W. D.; Kohler, M. A.; Campbell, J. P.; Godfrey, J. J. 2001. Phonetic, Idiolectal, and Acoustic Speaker Recognition, in *2001: A Speaker Odyssey. The Speaker Recognition Workshop, Crete, Greece*, 55–63.

Ariyaeeinia, A.; Sivakumaran, P. 1995. Effectiveness of orthogonal instantaneous and transitional feature parameters for speaker verification, in *Proc. IEEE Int. Conf. on Security Technology*, 79–84.

Atal, B. S. 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *J.A.S.A.* 55(6): 1304–1312.

Badran, E. F. M. F.; Selim, H. 2000. Speaker Recognition Using Artificial Neural Networks Based on Vowel phonemes, in *Proceedings of ICSP 2000, IEEE*, 2: 796–802.

Baum, L.; Petrie, T; Soules, G.; Weiss, N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Ann. Math Stat*. 41: 164–171.

Beek, B.; Neuberg, E. D.; Hodge, D. C. 1977. An assessment of the technology of automatic speech recognition for military applications, *IEEE Trans. Acoustics, Speech, Signal Processing* ASSP-25: 310–322.

Bimbot, F.; Blomberg, M.; Boves, L.; Genoud, D.; Hutter, H.-P.; Jaboulet, C.; Koolwaaij, J.; Lindberg, J.; Pierrot, J.-B. 2000. An overview of the CAVE project research activities in speaker verification. *Speech Communications* 31: 155–180.

Bricker, P. D.; Gnanadesikan, R.; Mathews, M. V.; Pruzansky, S.; Tukey, P. A. 1971. Statistical techniques for talker identification, *B.S.T.J.* 50: 1427–1454.

Campbell, J. P. 1997. Speaker Recognition: A Tutorial, in *Proceedings of the IEEE* 85(9): 1437–1462.

Campbell, J. P.; Reynolds, D.; Dunn, R. 2003. Fusing high- and low-level features for speaker recognition, in *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland*, 2665–2668.

Campbell, W. M.; Campbell, J. P.; Gleason, T. P.; Reynolds, D. A.; Shen, W. 2007. Speaker Verification Using Support Vector Machines and High-Level Features, *IEEE Transactions on Audio, Speech and Language Processing* 15(7): 2085–2094.

Chui, C. K. 1992. *Wavelets: a tutorial in theory and applications*. Boston: Academic Press.

Cooley, J. W.; Tukey, J. W. 1965. An algorithm for the machine calculation of complex Fourier series, *Mathematics of computation* 19(90): 297–301.

Daubechies, I. 1992. *Ten Lectures on Wavelets*, Philadelphia: SIAM.

Deller, J. R.; Hansen, J. H. L.; Proakis, J. G. 2000. Discrete-Time Processing of Speech Signals, *Piscataway (N.J.), IEEE Press*.

Dempster, A.; Laird, N.; Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Srar. Soc*. 39: 1–38.

Doddington, G. R. 1971. A method of speaker verification, *J.A.S.A*. 49(139) (A).

Doddington, G.; Liggett, W.; Martin, A.; Przybocki, M.; Reynolds, D. 1998 Sheeps, goats, lambs and wolves: a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation, in *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998)*, *Sydney, Australia.*

Duda, R.; Hart, P.; Stork, D. 2000. *Pattern Classification. Second ed*., New York: Wiley Interscience.

Endress, W.; Bambach, W.; Flosser, G. 1971. Voice spectrograms as a function of age, Voice Disguise and Voice Imitation, *J.A.S.A*. 49 6(2): 1842–1848.

Farooq, O.; Datta S. 2002. Mel-scaled wavelet filter based features for noisy unvoiced phoneme recognition, in *proc. of ICSLP 2002, Denver, Colorado, USA, Sept.* 16(20): 1017–1020.

Farrell, K. R.; Mammone, R. J.; Assaleh, K. T. 1994. Speaker Recognition Using Neural Networks and Conventional Classifiers, *IEEE Transactions on Speech and Audio Processing* 2(1): 194–205.

Fletcher, H. 1940. Auditory Patterns, *Rev.Mod.Phys* 12: 47–65.

Fredrickson, S. E.; Tarassenko, L. 1995. Text Independent Speaker Recognition using Neural Network Techniques. Artificial Neural Networks, *Conference Publication, IEEE* 409: 13–18.

Fukunaga, K. 1990. *Introduction to Statistical Pattern Recognition*. Second ed. London: Academic Press.

Furui, S. 2005. 50 years of progress in speech and speaker recognition, *proc. SPECOM, Patras, Greece*, 1–9.

Furui, S. 1974. An analysis of long-term variation of feature parameters of speech and its application to talker recognition, *Electronics and Communications in Japan* 57(A): 34–41.

Furui, S. 1981. Cepstral analysis technique for automatic speaker verification, *IEEE Transactions on Acoustics, Speech and Signal Processing* 29(2): 254–272.

Furui, S. 2001. *Digital Speech Processing, Synthesis and Recognition*, New York: Marcel Dekker.

Furui, S. 1997. Recent advances in speaker recognition. *Pattern Recognition Letters* 18(9): 859–872.

Furui, S.; Itakura, F.; Saito, S. 1972. Talker recognition by long time averaged speech spectrum, *Electronics and Communications in Japan* 55(A): 54–61.

Gersho, A.; Gray, R. 1991. *Vector Quantization and Signal Compression*, Boston: Kluwer Academic Publishers.

Gish, H.; Schmidt, M. 1994. Text-independent speaker identification, *IEEE Signal Processing Magazine* 11: 18–32.

Hansen, E. G.; Slyh, R. E.; Anderson, T. R. 2001. Formant and F0 Features for Speaker Recognition, in *2001: A Speaker Odyssey. The Speaker Recognition Workshop, Crete, Greece*, 25–30.

Hansen, E. G.; Slyh, R. E.; Anderson, T. R. 2004. Speaker Recognition using Phoneme-Specific GMMs, in *Odyssey 2004. The Speaker and Language Recognition Workshop. Toledo, Spain*.

Hansen, J.; Proakis, J. 2000. Discrete-Time Processing of Speech Signals, 2ed., *IEEE Press*, New York.

Hardt, D.; Fellbaum, K. 1997. Spectral subtraction and RASTA-filtering in text-dependent HMM-based speaker verification, in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1997), Munich, Germany*, 867–870.

Hattori, H. 1992. Text independent speaker recognition using neural networks, *Proceedings of IEEE*, 153–156.

He, J.; Liu, L.; Palm, G. 1999. A discriminative training algorithm for VQ-based speaker identification, *IEEE Trans. on Speech and Audio Processing* 7(3): 353–356.

Hebb, D. 1949. *Organization of Behavior*, New York: John Wiley & Sons.

Hermansky, H. 1994. RASTA processing of speech, *IEEE Trans. on Speech and Audio Processing* 2(4): 578–589.

Higgins, A.; Bahler, L. G.; Porter, J. E. 1991. Speaker verification using randomized phrase prompting, *Digital Signal Processing* 1: 89–106.

Huang, X.; Acero, A.; Hon, H.-W. 2001. *Spoken Language Processing: a Guide to Theory, Algorithm, and System Development*, New Jersey: Prentice-Hall.

Huggins, M.; Grieco, J. 2002. Confidence metrics for speaker identification, in *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002). Denver, Colorado, USA*, 1381–1384.

Hui-Ling, L. 2002. *Toward a high-quality singing synthesizer with vocal texture control*: PhD thesis, Stanford University.

Hume, J. 1997. Wavelet-like regression features in the cepstral domain for speaker recognition in *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997), Rhodos, Greece*, 2339–2342.

Jawerth, B.; Sweldens, W. 1994. An overview of wavelet based multiresolution analyses, *SIAM Review* 36: 377–412.

Juang, B.-H., *et al*. 1987. A vector quantization approach to speaker recognition, *AT & T Technical Journal* 66: 14–26.

Kabal P.; Ramachandran, R. P. 1986. The Computation of Line Spectral Frequencies Using Chebyshev Polynomials, *IEEE Transactions on Acoustic, Speech, and Signal Processing* ASSP-34(6): 1419–1426.

Kerstholt, J.; Jansen, E.; Van Amelsvoort, A.; Broeders, A. 2003. Earwitness line-ups: effects of speech duration, retention interval and acoustic environment on identification accuracy, in *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland*, 709–712.

Kinnunen, T. 2005. *Optimizing Spectral Feature Based Text-Independent Speaker Recognition*: Academic dissertation, University of Joensuu, 8–9.

Kinnunen, T., *et al*. 2000. Comparison of Clustering Algorithms in Speaker Identification, in *Proc. IASTED Int. Conf. Signal Processing and Communications (SPC 2000), Marbella, Spain*, 222–227.

Li, Q.; Juang, B.-H.; Lee, C.-H. 2000. Automatic verbal information verification for user authentication, *IEEE Trans. on Speech and Audio Processing* 8: 585–596.

Li, X.; Mak, M.; Kung, S. 2001. Robust speaker verification over the telephone by feature recuperation, in *Proc. 2001 Int. Symposium on Intelligent Multimedia, Video, and Speech Processing, Hong Kong*, 433–436.

Linde, Y.; Buzo, A.; Gray, R. 1980. An algorithm for vector quantizer design, *IEEE Transactions on Communications* 28(1): 84–95.

Lipeika, A. 2000. Segmentation of random sequences, *Informatica* 11(3): 243–256.

Lipeika, A.; Lipeikienė, J. 1995. Speaker Identification using Vector Quantization, *Informatica* 6(1): 167–180. ISSN 0868-4952.

Liu, L.; He, J.; Palm, G. 1997. A comparison of human and machine in speaker recognition, in *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997), Rhodos, Greece*, 2327–2330.

Long, C. J.; Datta, S. 1996. Wavelet based feature extraction for phoneme recognition, in *Proc. of 4th Int. Conf. of Spoken Language Processing, Philadelphia, USA* 1: 264–267.

Majewski, W.; Mazur-Majewska, G. 1999. Speech signal parametrization for speaker recognition under voice disguise conditions, in *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 1999), Budapest, Hungary*, 1227–1230.

Makhoul, J. 1975. Linear prediction: a tutorial review, in *Proceedings of IEEE* 63(4): 561–580.

Mallat, S. G. 1998. *A Wavelet Tour of Signal Processing*, New York: Academic Press,.

Mammone, R.; Zhang, X.; Ramachandran, R. 1996. Robust speaker recognition: a feature based approach, *IEEE Signal Processing Magazine* 13(5): 58–71.

Markel, J.D.; Gray, A. H. 1976. *Linear prediction of speech*, Berlin: Springer-Verlag. ISBN 3-540-07563-1.

Martin, A.; Doddington, G.; Kamm, T.; Ordowski, M.; Przybocki, M. 1997. The DET Curve in Assessment of Detection Task Performance, in *Eurospeech'97*, *Rhodes, Greece*, 1895–1898.

Mary, L.; Murty, K. S. R; Prasanna, S. R. M; Yegnanarayana, B. 2004. Features for Speaker and Language Identification, in *Odyssey 2004. The Speaker and Language Recognition Workshop*. *Toledo, Spain*.

Matsui, T.; Furui, S. 1992. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs, in *Proc. ICSLP*, II,157–160.

Matsui, T.; Furui, S. 1993. Concatenated phoneme models for text-variable speaker recognition, in *Proc. ICASSP*, II, 391–394.

Matsui, T.; Furui, S. 1994. Similarity normalization method for speaker verification based on a posteriori probability, in *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, 59–62.

McLachlan, G. 1988. *Mixture Models*, New York: Marcel Dekker.

Naik, J. M. 1990. Speaker Verification: A Tutorial, *IEEE Communications Magazine* 28(1): 42–48.

Naik, J. M.; Netsch, L. P.; Doddington, G. R. 1989. Speaker verification over long distance telephone lines, in *Proc. ICASSP* 1: 524–527.

Navakauskas, D. 2000. *Skaitmeninio signalų apdorojimo priemonės. Dirbtinių neuronų tinklai*, Vilnius: Technika.

Ning, D.; Chandran, V. 2004. The Effectiveness of Higher Order Spectral Phase Features in Speaker Identification, in *Odyssey 2004. The Speaker and Language Recognition Workshop*. *Toledo, Spain.*

Obrecht, R. A. 1988. A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals, *IEEE Transactions on Acoustics. Speech and Signal Processing* 36(1): 29–40.

Ong, S.; Sridharan, S.; Yang, C.-H.; Moody, M. P. 1996. Comparison of Four Distance Measures for Long Time Text-Independent Speaker Identification, *ISSPA*, 369–372.

Oppenheim, A. V.; Schafer, R. W.; Buck, J. R. 1999. Discrete-Time Signal Processing, 2nd ed., *Upper Saddle River*, New York: Prentice Hall.

Ore, B. M; Slyh, R. E; Hansen, E. G. 2006. Speaker Segmentation and Clustering using Gender Information, *2006 IEEE Odyssey, The Speaker and Language Recognition Workshop, San Juan, Puerto Rico*, 1–8.

Orsag, F. 2004. *Biometric Security Systems, Speaker Recognition Technology*: Dissertation, BRNO university of technology.

Osuna, E.; Freund, R.; Girosi, F. 1997. Training Support Vector Machines: An application to face detection, in *Proceedings of CVPR'97*, *Puerto Rico*, 130–136.

Phythian, M.; Ingram, J.; Sridharan, S. 1997. Effects of speech coding on text-dependent speaker recognition, in *Proceedings of IEEE Region 10 Annual Conference on Speech and Image Technologies for Computing and Telecommunications* (TENCON'97), 137–140.

Poritz, A. B. 1982. Linear predictive hidden Markov models and the speech signal, in *Proc. ICASSP* 2: 1291–1294.

Prabhakar, S.; Pankanti, S.; Jain, A. 2003. Biometric recognition: security and privacy concerns, *IEEE Security & Privacy Magazine* 1: 33–42.

Pruzansky, S. 1963. Pattern-matching procedure for automatic talker recognition *J.A.S.A*. 35: 354–358.

Quatieri, T.; Reynolds, D.; O'leary, G. 2000. Estimation of handset nonlinearity with application to speaker recognition, *IEEE Trans. on Speech and Audio Processing* 8(5): 567–584.

Rabiner, L. R,; Cheng, M. J.; Rosenberg, A. E.; Mcgonegal, C. A. 1976. A Comparative Performance Study of Several Pitch Detection Algorithms, *IEEE Transactions on Acoustics, Speech and Signal Processing* ASSP-24(5): 399–418.

Rabiner, L; Juang, B. H. 1986. An introduction to hidden Markov models, *IEEE ASSP Mag*. 3(1): 4–16.

Rabiner, R. L.; Rosenberg, A. E.; Levinson, S. E. 1978. Considerations in dynamic time warping algorithms for discrete word recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing* 26(5): 575–582.

Rabiner, L.R.; Schafer, R.W. 1978. *Digital processing of speech signals*, New Jersey: Prentice Hall, ISBN 0-13-213603-1.

Reynolds, D. 2002. An Overview of Automatic Speaker Recognition Technology, in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, *Orlando, Florida, USA,* 4072–4075.

Reynolds, D. 1994. Experimental evaluation of features for robust speaker identification, *IEEE Trans. on Speech and Audio Processing* 2(4): 639–643.

Reynolds, D.; Andrews, W.; Campbell, J.; Navratil, J.; Peskin, B.; Adami, A.; Jin, Q.; Klusacek, D.; Abramson, J.; Mihaescu, R.; Godfrey, J.; Jones, D.; Xiang, B. 2003. The SuperSID project: exploiting high-level information for high-accuracy speaker recognition, in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, *Hong Kong*, 784–787.

Reynolds, D.; Rose, R. 1995. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models, *IEEE transactions on speech and audio processing* 3(1): 72–83.

Rodman, D. R. 1999. *Computer Speech Technology*, Boston, Mass.: Artech House.

Rose, P. 2002. *Forensic Speaker Identification*, London: Taylor & Francis.

Rose, R.; Reynolds, R. A. 1990. Text independent speaker identification using automatic acoustic segmentation, in *Proc. ICASSP*, 293–296.

Rosenberg, A. E.; Sambur, M. R. 1975. New techniques for automatic speaker verification, *IEEE Trans. Acoustics, Speech, Signal Proc*. ASSP-23(2): 169–176.

Rosenblatt, F. 1957. *The perceptron: A perceiving and recognizing automaton (project PARA)*, Technical Report 85-460-1, Cornell Aeronautical Laboratory.

Sakoe, H. 1979. Two-level DP-matching – A dynamic programming-based pattern matching algorithm for connected word recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing* 27(6): 588–595.

Sakoe, H.; Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing* 26(1): 43–49.

Salna, B.; Mambro, G. D. 2006. *Method and System for Bio-metric Voice Print Authentication*, Patent USA, Nr. EP-0001915294.

Sambur, M. R. 1972. *Speaker recognition and verification using linear prediction analysis*: Ph. D. Dissert., M.I.T.

Sarikaya, R.; Hansen, H. L. 2000. High resolution speech feature parameterization for monophone-based stressed speech recognition, *IEEE Signal Processing Letters* 7(7): 182–185.

Sarikaya, R.; Pellom, B.L.; Hansen H. L. 1998. Wavelet packet transform features with application to speaker identification, in *Proc. of IEEE Nordic Signal Processing Symp*., *Visgo, Denmark*, 81–84.

Schmidt-Nielsen, A.; Crystal, T. 2000. Speaker verification by human listeners: experiments comparing human and machine performance using the nist 1998 speaker evaluation data, *Digital Signal Processing* 10: 249–266.

Siafarikas, M.; Ganchev, T.; Fakotakis, N. 2004. Wavelet Packet Based Speaker Verification, in *Odyssey 2004. The Speaker and Language Recognition Workshop*. *Toledo, Spain*.

Simpson, P. K. 1990. *Artificial Neural Systems: Foundations, Paradigms, Applications, And Implementations*, New York: Pergamon Press.

Slyh, R. E.; Hansen, E. G.; Anderson, T. R. 2004. Glottal Modeling and Closed-Phase Analysis for Speaker Recognition, in *Odyssey 2004. The Speaker and Language Recognition Workshop*. *Toledo, Spain*, 315–322.

Soong, F. K.; Rosenberg, A. E.; Juang, B-H. 1987. A vector quantization approach to speaker recognition, *AT&T Technical Journal* 66: 14–26.

Tamulevičius, G. 2008. *Pavienių žodžių atpažinimo sistemų kūrimas*: daktaro disertacija, Vilniaus Gedimino technikos universitetas ir Matematikos ir informatikos institutas. Vilnius: Technika. 124 p.

Tishby, N. 1991. On the application of mixture AR hidden Markov models to text independent speaker recognition, *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-30(3): 563–570.

Toledo-Ronen, O. 2001. Speech Detection for Text-Dependent Speaker Verification, in *2001: A Speaker Odyssey. The Speaker Recognition Workshop*, *Crete, Greece*.

Tufekci, Z.; Gowdy, J. N. 2000. Feature extraction using discrete wavelet transform for speech recognition, in *Proc. of IEEE*, Southeastcon 2000, 116–123.

Vapnick, V. 1995. *The Nature of Statistical learning theory*, New York: Springer-Verlag.

Weber, K, *et al*. 2002. Evaluation of Formant-Like Features for ASR, in *ICSLP-2002*, 2101–2104.

White, G. M.; Neely, R. B. 1976. Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming, *IEEE Transactions on Acoustics, Speech and Signal Processing* 24(2): 183–188.

Wong, L. P.; Russell, M. J. 2001. Speaker Verification Under Additive Noise Conditions with Non-stationary SNR Using PMC, in *2001: A Speaker Odyssey. The Speaker Recognition Workshop*, *Crete, Greece*, 95–100.

Zelinski, R.; Class, F. 1983. A Segmentation Algorithm for Connected Word Recognition Based on Estimation Principles, *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-31(4): 818–827.

Zilca, R. D.; Pelecanos, J. W.; Chaudhari, U. V.; Ramaswamy, G. N. 2004. Real Time Robust Speech Detection for Text Independent Speaker Recognition, in *Proceedings of Odyssey-04, The Speaker and Language Recognition Workshop*, *Toledo, Spain*, 123–128.

Zilovic, M. S.; Ramachandran, R. P.; Mammone, R. 1998. Speaker Identification Based on the Use of Robust Cepstral Features Obtained from Pole – Zero Transfer Functions, *IEEE Transactions on Speech and Audio Processing* 6(3): 260–268.

Autoriaus mokslinių publikacijų disertacijos tema sąrašas

Straipsniai recenzuojamuose mokslo žurnaluose

Kamarauskas, J. 2006. Automatic Segmentation of the Phonemes using Artificial Neural Networks, *Electronics and Electrical Engineering* 8(72): 39–42. ISSN 1392–1215.

Kamarauskas, J. 2008. Speaker recognition using Gaussian Mixture Models, *Electronics and Electrical Engineering* 5(85): 29–32. ISSN 1392–1215 (ISI Master Journal List)

Straipsniai kituose leidiniuose

Kamarauskas, J. 2007. Kalbančiojo atpažinimas taikant vektorinį kvantavimą [CD-ROM], in *Informacinės technologijos 2007*, 42–46. ISSN 1822-6337.

Šalna, B.; Kamarauskas, J. 2005. Automatinio asmens atpažinimo iš balso problemos ir perspektyvos kriminalistikoje, *Jurisprudencija* 66(58): 140–145. ISSN 1392-6195.

Juozas KAMARAUSKAS

ASMENS ATPAŽINIMAS PAGAL BALSĄ

Daktaro disertacija

Technologijos mokslai,
informatikos inžinerija (07T)

SPEAKER RECOGNITION BY VOICE

Doctoral Dissertation

Technological Sciences,
Informatics Engineering (07T)

Autoriaus kontaktiniai duomenys
juozas.kamarauskas@gmail.com

2009 04 16. 14,25 sp. l. Tiražas 20 egz.
Vilniaus Gedimino technikos universiteto leidykla „Technika“,
Saulėtekio al. 11, 10223 Vilnius,
http://leidykla.vgtu.lt
Spausdino UAB „Biznio mašinų kompanija“,
J. Jasinskio g. 16A, 01112 Vilnius
http://www.bmk.lt