Automaattisen puhujanvarmennuksen päätöslogiikkaa

Teemu Kilpeläinen, Op.Nro.:136350

17.9.2001

Kandidaatintutkielma

Joensuun yliopisto

Tietojenkäsittelytieteen laitos

Tiivistelmä

Tässä tutkielmassa käydään läpi automaattisten puhujanvarmennusjärjestelmien toimintaan liittyvää *päätöslogiikkaa*. Toimiva päätöslogiikka on kiinteä osa korkean luotettavuuden ja käytettävyyden omaavaa varmennusjärjestelmää. Lisäksi tutkielmassa tarkastellaan varmennusjärjestelmien arviointiin soveltuvia työkaluja, joita ovat mm. järjestelmän keskimääräistä virhealttiutta mittaavat mittarit *FR* (False Rejection rate), *FA* (False Appentance rate) ja *EER* (Equal Error Rate). Näitä samoja työkaluja käytetään myös järjestelmän *kynnysarvojen* suunnitteluun. Päätöslogiikan suunnitteluun perehdytään selvittämällä varmennuspäätöksen peruskäsitteitä, sekä antamalla esimerkkejä puhujille asetettavan kynnysarvon suunnitteluun tarkoitetuista metodeista.

SISÄLLYSLUETTELO

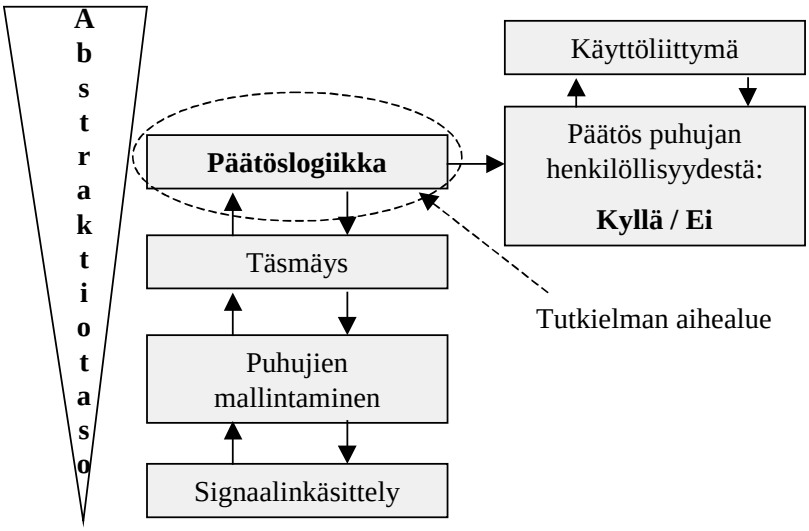
1 Johdanto

Ihmisen henkilöllisyyden varmentaminen kiinteä osa nykyihmisen arkipäivää. Henkilöllisyys voidaan varmentaa monella eri tavalla, joista ehkä yleisimpiä ovat erilaiset tunnussanat ja PIN (*Person Identification Numbers*)-koodit. Näiden henkilöllisyydenvarmennusmenetelmien heikkoutena on niiden riippumattomuus henkilön omista henkilökohtaisista ominaisuuksista [5]. Tunnussanat ja PIN-koodit ovat itse asiassa ainoastaan merkkijonoja jotka voivat helpostikin joutua sivullisen käsiin ja antaa hänelle näin mahdollisuuden esiintyä tunnusluvun omistajana esimerkiksi pankkiautomaatilla. Tämä on vakava turvallisuusuhka, jonka torjumiseen tuovat apua ihmisen omiin henkilökohtaisiin ominaisuuksiin pohjautuvat, *biometriset*, henkilöllisyydenvarmennusmenetelmät [1]. Tällaisia varmennusmenetelmiä ovat esimerkiksi sormenjälkien-, silmän rakenteen- ja puhujan henkilöllisyyden varmentamiseen puheäänestä liittyvät menetelmät [3, 5]. Tässä tutkielmassa keskitytään puhujanvarmennusmenetelmän kuvaamiseen ja erityisesti kyseiseen menetelmään kuuluvan päätöslogiikan tarkasteluun. Puhujan varmennukseen pohjautuvia käyttäjän henkilöllisyydenvarmentamismenetelmiä käytetään yleensä osana laajempaa turvallisuusjärjestelmää, jossa käytetään rinnakkain muitakin varmennusmenetelmiä, kuten tavallinen metalliavain tai näppäiltävä PIN-koodi. Tällaisilla, useaan eri tunnistusmetodiin pohjautuvilla järjestelmillä saavutetaan niin sanottu *vahva tunnistaminen* [8].

Tutkielmassa käydään läpi tarkemmin automaattiseen puhujanvarmennusjärjestelmien evaluointiin liittyviä osatekijöitä. Näitä ovat mm. järjestelmän arviointimittarit *EER* (Equal Error Rate) ja *ROC* (Receiver Operating Characteristics). Lisäksi tutkielmassa keskitytään selvittämään varmennusjärjestelmien sisältämää *päätöslogiikkaa* (desicion logic), joka on oleellinen osa toimivaa ja luotettavaa järjestelmää.

Automaattinen puhujanvarmennus voidaan jakaa yleisellä tasolla osatehtäviin (Kuva 1.): järjestelmän alimmalla abstraktiotasolla sijaitsevat puhenäytteen käsittelyyn tarvittavat signaalinkäsittelyoperaatiot. Seuraavaa tasoa kutsutaan puhujan mallinnukseksi. Tässä vaiheessa järjestelmä luo signaalinkäsittelyn tuloksena syntyneestä tietojoukosta puhujamallin, joka kuvaa puhujan äänen yksilöllisiä piirteitä. Täsmäysvaiheessa järjestelmä vertailee puhujan antamaa ääninäytettä

aiemmin tietokantaan tallennettuun kyseisen puhujan ääninäytteeseen. Seuraavalla tasolla ylöspäin mentäessä sijaitsee järjestelmän päätöslogiikka, jossa järjestelmä tekee päätöksen puhujan hyväksymisestä tai hylkäämisestä. Ylimmällä tasolla on käyttöliittymä, jonka välityksellä käyttäjä on yhteydessä varmennusjärjestelmään.



Kuva 1, Puhujanvarmennusjärjestelmä yleisellä tasolla.

Tämän tutkielman loppuosa jakautuu luvuittain seuraavasti: luvussa kaksi luodaan yleiskatsaus automaattiseen puhujantunnistukseen, kolmas luku käsittelee puhujanvarmennuksen evaluointia käyden läpi varmennusjärjestelmien arviointimittareita, neljännessä luvussa perehdytään varmennusjärjestelmien päätöslogiikan suunnitteluun ja viides luku on yhteenveto tutkielman sisällöstä.

2 Puhujantunnistuksen periaatteita

Puhujantunnistustehtävä voidaan ajatella yksinkertaisesti prosessina, jossa tietokone tunnistaa puhujan henkilöllisyyden tämän äänen ominaispiirteiden perusteella [12]. Tässä luvussa selvitetään puhujantunnistustehtävän perusteita ja kyseiseen tehtävään liittyviä ongelmakenttiä yleisellä tasolla. Lisäksi tarkastellaan puhujanvarmennusjärjestelmien vahvuuksia ja heikkouksia suhteessa muihin henkilöntunnistusjärjestelmiin.

Automaattinen puhujantunnistus voidaan jakaa tehtävänä muiden biometristen tunnistusmenetelmien tapaan kahteen osaan [3, 6, 7, 8]:

1. Puhujan *identifiointiin* (Speaker Identification), jossa järjestelmän tavoitteena on erottaa puhuja rekisteröityneiden puhujien joukosta.
2. Puhujan *varmentamiseen* (Speaker Verification), jossa puhujantunnistusjärjestelmän tehtävänä on varmentaa käyttäjän antama väite henkilöllisyydestään tai hylätä väite.

Yleiskielessä käytetään usein termiä *puhujantunnistus* yleisnimenä automaattisille puhujanvarmennus- ja puhujantunnistamistehtäville. Tämä tutkielma keskittyy erityisesti varmennustehtävään liittyvän ongelmakentän selvittämiseen. Sekä puhujan varmentaminen että puhujan identifiointi voidaan jakaa karkeasti kahteen toisistaan erotettavaan suoritusvaiheeseen; *opetusvaiheeseen* ja *tunnistusvaiheeseen* [3, 6]. Näitä käsitellään lyhyesti seuraavassa.

Opetusvaihe

Opetusvaiheen tarkoituksena on opettaa järjestelmälle kunkin puhujan *puhujamalli,* eli ”äänijälki”, jonka avulla kyseinen puhuja voidaan hakea myöhemmin tietokannasta [8]. Puhujamalli sisältää paljon informaatiota puhujan yksilöllisistä puheominaisuuksista. Tätä informaatiota kutsutaan usein puheen *piirteiksi* ja suorituksen osavaihetta, jossa puhenäytteestä etsitään piirteitä, kutsutaan *piirreirrotukseksi*. Piirreirrotusta varten tuotettu analoginen puhesignaali muutetaan ensin analogisesta digitaaliseksi eli tietokoneen ymmärtämään muotoon [4].

Puhujamallin on tarkoituksena kuvata puhujan äänisignaalin vaihteluita mahdollisimman hyvin, jotta malli olisi toimiva myös tilanteissa, joissa puhuja on vilustunut tai hänen puheäänensä on jotenkin muulla tavalla muuttunut. Puhujamalleja voidaan luoda myös useita kappaleita jokaista käyttäjää kohti, tällä tavalla puheäänen mahdollinen muuttuminen voidaan huomioida vielä paremmin. Puhujamallit voidaan jakaa yleisesti kahteen eri luokkaan [8]: *sapluunamalleihin* (template models) ja *stokastisiin* malleihin (stochastic models). Sapluunamallien tarkoituksena on pyrkiä mallintamaan tiettyä jotakin lausahdusta useasta piirrevektorista muodostetun sarjan pohjalta rakennetun keskiarvosarjan perusteella [biometrics.org]. Stokastisissa malleissa puheentuottamisprosessi oletetaan satunnaisprosessiksi, johon tarvittavat parametrit voidaan arvioida tarkasti jollain määritellyllä menetelmällä [8]. Tärkeä tehty tutkimustulos on, että piirreirrotus parantaa tehokkuuden lisäksi myös puhujantunnistamistehtävän suorittavan järjestelmän tarkkuutta, koska piirreirrotus poistaa äänisignaalista tunnistuksessa turhaa tietoa [6]. Turhaa, piirreirrotuksessa poistettavaa tietoa ovat esimerkiksi ääninäytteessä olevat tauot ja taustamelu. Piirreirrotusvaiheessa syntyneistä piirrevektoreista luodaan puhujamalli, joka tallennetaan järjestelmän tietokantaan. Opetusvaihe on usein samanlainen sekä puhujan identifiointi-, että puhujanvarmennustehtävissä.

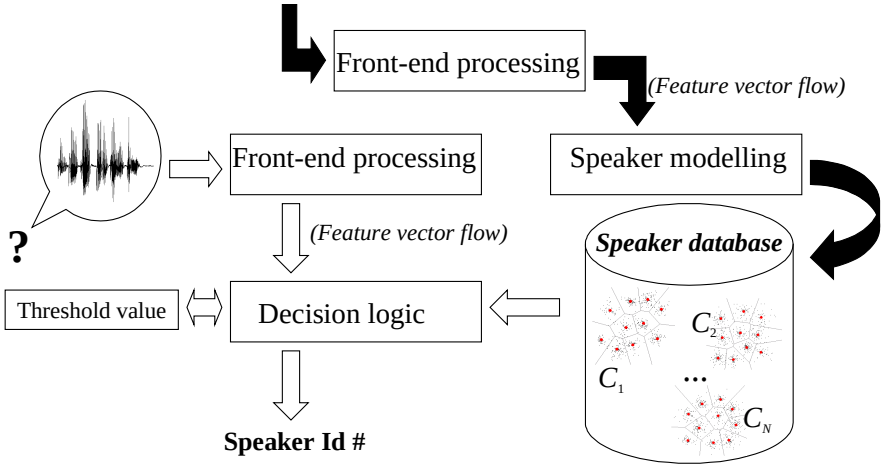
Tunnistusvaihe

Tunnistusvaiheessa toteutetaan varsinainen puhujantunnistus. Puhuja antaa tunnistusjärjestelmälle ensin ääninäytteensä, jolle tehdään opetusvaiheen mukaisesti piirreirrotus. Piirreirrotuksen tuloksena syntynyttä puhujamallia verrataan tämän jälkeen järjestelmän tietokannassa oleviin puhujamalliin/puhujamalleihin. Tunnistustehtävä voi olla joko *avoin joukko* (open set), jossa puhujan puhujamallia ei välttämättä ole olemassa järjestelmän tietokannassa, tai *suljettu joukko* (closed set), jolloin puhujan malli on olemassa [3, 6]. Tunnistusvaiheet eroavat toisistaan puhujan identifiointi- ja varmennustehtävässä.

2.1 Identifiointitehtävä

Jos halutaan tietää, ketä puhujatietokannan puhujaa tuntematon puhuja eniten muistuttaa, on kyseessä puhujan identifiointitehtävä (Kuva 2.). Identifiointi tehdään suljetun joukon tapauksessa

laskemalla todennäköisyys puhujamallin sisältämien *piirrevektoreiden* ja kaikkien järjestelmän tietokannassa olevien puhujamallien suhteen ja tunnistetaan puhuja sen mukaan, jonka puhujamallit täsmäävät parhaiten. Avoimen joukon tapauksessa puhuja tunnistetaan, jos löydetään *kynnysarvon* (threshold value) alittava eroavaisuus kyseisen puhujan ja jonkin tietokannassa olevan puhujamallin kanssa. Kynnysarvon alittavista puhujamalleista valitaan kaikkein parhaiten täsmäävä.

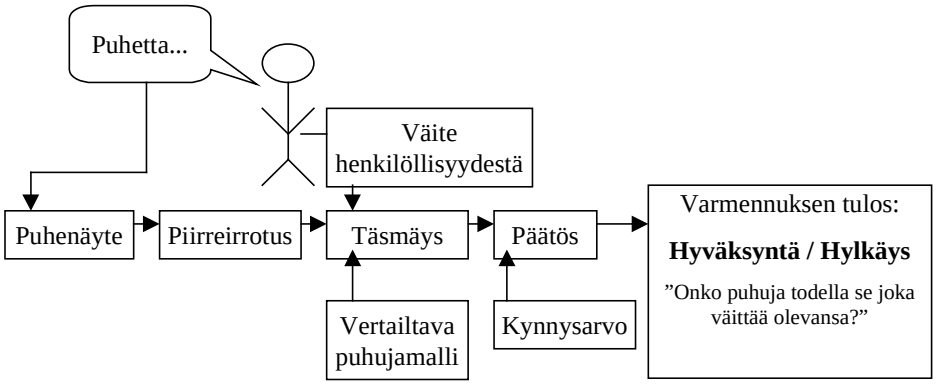


Kuva 2, Puhujan identifiointi kaaviokuvana.

2.2 Varmennustehtävä

Puhujanvarmennusjärjestelmän tarkoituksena on varmistaa, onko puhuja se joka hän väittää olevansa. *Väite henkilöllisyydestä* (identity claim) on joko tosi tai epätosi. Varmennustehtävässä verrataan puhujan piirrevektoreita ainoastaan väitetyn puhujan puhujamalliin, toisin kuin identifiointitehtävässä, jossa vertailu tehdään kaikkien puhujamallien välillä. Jos vertailussa päästään *kynnysarvoa* (threshold value) pienempään eroavuuteen, annetaan päätös ”henkilö oli kysytty henkilö”, muuten ”henkilö ei ole kysytty henkilö”. Jos puhuja ei ole kysytty henkilö, voi tunnistusalgoritmi näyttää kielteisen päätöksen lisäksi pisteytyksen jonka perusteella varmennus epäonnistui. Esimerkiksi: ”löydettiin yli 2 prosentin eroavaisuus, puhuja ei ole kysytty henkilö”. Tässä tutkielmassa käydään läpi kynnysarvon määrittelemistä tarkemmin luvussa 4. Turvallisuustehtävässä toimiva varmennusjärjestelmä voi varmennuksen hylkäämisen lisäksi ilmoittaa hylkäyksestä suoraan esimerkiksi vartiointiliikkeeseen. Varmennustehtävä on *binäärinen luokit-*

teluongelma, käyttäjän antama väite henkilöllisyydestään joko hyväksytään tai hylätään [8] (Kuva 3.).



Kuva 3, Puhujanvarmennus kaaviokuvana.

2.3 Sisällöstä riippuvat ja sisällöstä riippumattomat tunnistustehtävät

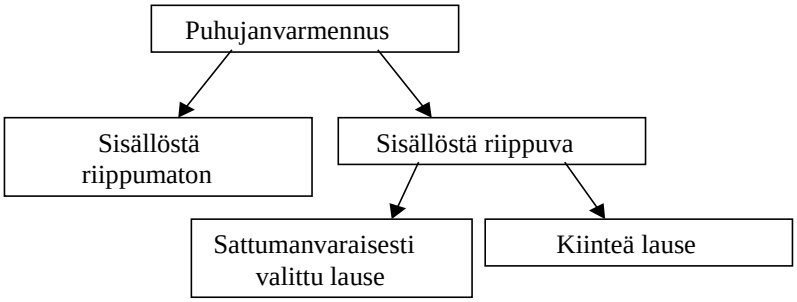
Puhujanvarmennus- ja puhujan identifiointitehtävät voivat olla [3, 6]:

1. *Tekstin sisällöstä riippuvia* (text-dependent) tehtäviä, joissa käyttäjä antaa kirjautumisvaiheessa puhenäytteenään joitakin opetusvaiheessa antamistaan sanoista. Tunnistukseen käytettävä(t) lause(et) on siis jo valmiiksi järjestelmän tiedossa.
2. *Tekstin sisällöstä riippumattomia* (text-independent) tehtäviä, joissa käyttäjän kirjautumisvaiheessa antama puhenäyte ei riipu opetusvaiheessa annetusta näytteestä, vaan voi sisältää muitakin sanoja.

Sisällöstä riippuva tunnistustehtävä

Käyttäjä antaa varmennusjärjestelmälle järjestelmän opetusvaiheessa ääninäytteensä, jonka sisältö on ennalta määrätty ja koostuu esimerkiksi numeroista ja sanoista. Järjestelmä luo tästä ääninäytteestä puhujamallin tälle kyseiselle puhujalle. Tämä puhujamalli toimii jatkossa kyseisen käyttäjän ”avaimena” järjestelmään. Kun käyttäjä yrittää myöhemmin kirjautua sisään järjestelmään, järjestelmä vaatii käyttäjää lukemaan opetusvaiheessa käytettyjä numeroita ja sanoja varmentamista varten. Näytteen saatuaan järjestelmä vertaa käyttäjän ääninäytettä ennalta tallennet-

tuun puhujamalliin. Sisällöstä riippuvan varmennusjärjestelmän vaatimat lauseet voivat olla *kiinteitä* (fixed phrase), tai *sattumanvaraisesti valittuja* (promted phrase) [7] (Kuva 4.). Kiinteitä lauseita käytettävissä järjestelmissä käyttäjältä vaaditaan jokaisella sisäänkirjautumiskerralla sama ääninäyte, kun taas jälkimmäisellä tavalla toimiva järjestelmä arpoo jokaisella sisäänkirjautumiskerralla käyttäjälle luettavaksi uuden sisällön omaavan tekstin. Sattumanvaraisella sisällöllä ja tiukalla aikamarginaalilla ääninäytteitä vaativat järjestelmät ovat kiinteitä puhenäytteitä vaativia järjestelmiä turvallisempia, koska näillä tekniikoilla voidaan estää, tai ainakin huomattavasti vaikeuttaa tallennetun puheäänen käyttömahdollisuutta järjestelmän sisäänkirjautumisessa [8]. Toisaalta järjestelmän käytettävyyden kannalta toistettavan lauseen satunnaistaminen saattaa olla epäloogista ja käyttäjälle jopa epämiellyttävää: toistettavana lauseena voisi esimerkiksi olla: ”viisi kaksi ankka yksi koira yhdeksän pää…”.



Kuva 4, Puhujanvarmennuksen luokittelu sisältöriippuvuuden mukaan.

Tekstin sisällöstä riippuvainen puhujanvarmennusjärjestelmä sopii hyvin esimerkiksi teollisuuslaitoksen ovenavausjärjestelmään, jossa käyttäjäryhmä on ennalta määrätty ja kaikilta käyttäjiltä on mahdollista saada opetusvaihetta varten ääninäytteet. Lisäksi puhenäytteen ei tarvitse sisäänkirjautumisvaiheessakaan olla kuin korkeintaan opetusvaiheessa käytetyn ääninäytteen mittainen [3, 5].

Sisällöstä riippumaton tunnistustehtävä

Tekstin sisällöstä riippumaton puhujanvarmennusjärjestelmä ei vaadi käyttäjältä tiettyä, ennalta määrätyn sisällön omaavaa ääninäytettä varmentamista varten. Varmennettavan henkilön ääninäyte voidaan ottaa järjestelmän opetusvaihetta varten jo olemassa olevasta käyttäjän ääninäytteestä. Järjestelmän opetukseen tarkoitetun ääninäytteen sisällöllä ei ole varsinaista sisältövaatimusta, mutta näyte on sitä parempi mitä monipuolisemmin se sisältää käytettävän kielen yleisim-

piä äänteitä [3]. Monipuolinen opetusääninäyte parantaa varmennustehtävän tarkkuutta, koska tunnistusvaiheessa käytettävä ääninäyte ei tällaisessa tilanteessa useinkaan sisällä kokonaisia samoja sanoja kuin järjestelmän opetusvaiheessa käytettiin.

Tekstin sisällöstä riippumaton puhujanvarmennusjärjestelmä sopii hyvin esimerkiksi poliisin käyttöön puhelinkuuntelutehtävän yhteydessä. Poliisin tutkija käyttää tällöin järjestelmän opettamiseen hallussaan olevaa epäillyn aiempaa ääninäytettä ja varmentaa puhujan henkilöllisyyden tuoreemmasta ääninäytteestä [9].

Tekstin sisällöstä riippumatonta puhujanvarmennusjärjestelmää ei kuitenkaan saada yhtä tarkaksi kuin tekstin sisällöstä riippuvaa järjestelmää [5, 8]. Lisäksi sisällöstä riippumaton järjestelmä tarvitsee tunnistusvaihetta varten jopa 10-30 sekunnin mittaisen ääninäytteen [5].

3 Varmennusjärjestelmän arviointi

Puhujanvarmennusjärjestelmien *arvioinnin* (evaluation) päämääränä ovat entistä varmemmat ja luotettavammat varmennusjärjestelmät. Varmennusjärjestelmien luotettavuutta arvioidaan erilaisilla tavoilla, joista yleisimmin käytettyjä ovat järjestelmän virheherkkyyttä mittaavat virhetulosparametrit:

1. *FR* (False Rejection Rate), joka kertoo kuinka monta prosenttia aidoista puhujista järjestelmä keskimäärin hylkää.
2. *FA* (False Acceptance Rate), joka kertoo kuinka monta prosenttia huijareista pääsee keskimäärin sisään järjestelmään.

Virhetulosparametrien kuvaajat voidaan yhdistää samaan koordinaatistoon ns. ROC (Receiver Operating Characteristics) –kuvaajaan, josta FR- ja FA-kuvaajien leikkauspiste kertoo järjestelmän *keskimääräisen virhealttiuden* (EER, Equal Error Rate). Lisäksi varmennusjärjestelmiä voidaan vertailla järjestelmän *kokonaissuoritusaikojen* (VT, Verification Throughput) perusteella. Näitä käsitellään seuraavassa lyhyesti.

3.1 FR (False Rejection Rate)

FR-tyypin virhe tapahtuu kun järjestelmä hylkää aidon puhujan. Monissa tunnistusjärjestelmissä tällaisille virheille annetaan alle 1 prosentin marginaali kaikista varmennetuista ääninäytteistä [5]. Esimerkiksi erittäin laajan käyttäjäkunnan omaavalla puhelinvälitteisellä varmennusjärjestelmällä jossa käyttäjiä on jopa miljoonia, jo yli 1 prosentin virheosuus voisi olla joustavan käytön kannalta liian suuri [5]. Virhe johtuu usein järjestelmän lisäksi myös ääninäytteen huonosta teknisestä laadusta. Puhelinvälitteisessä järjestelmässä puhelinlinjasta johtuvat rasahdukset ja äänen katkokset hankaloittavat puhujan varmennusjärjestelmän toimintaa. Virheestä ei koidu yleensä vakavaa tietoturva- tms. vahinkoa ja virhe voidaan usein poistaa käsittelemällä käyttäjän uusi ääninäyte.

3.2 FA (False Acceptance Rate)

FA-tyypin virhe tapahtuu kun järjestelmä hyväksyy huijarin (impostor). Nämä virheet ovat huomattavasti *FR* –tyyppisiä virheitä vakavampia. Virhe käytännössä romahduttaa käytettävän järjestelmän luotettavuuden esimerkiksi ovenaukaisupalvelimena muuten tarkasti varjellussa teollisuuskohteessa. Yleisesti hyväksytty virhemarginaali tämän tyyppisten virheiden esiintymisille tarkasti suojelluissa järjestelmissä on 0.1 prosenttia kaikista varmennetuista ääninäytteistä [5]. Erittäin tarkasti suojelluissa kohteissa joissa on vähän käyttäjiä (n. 50 – 1000 kpl.), kuten armeijan tietokonekeskuksessa, varmennusjärjestelmän virhemarginaalin *FA*-tyyppisten virheiden osalta tulee olla jopa tätäkin pienempi [5].

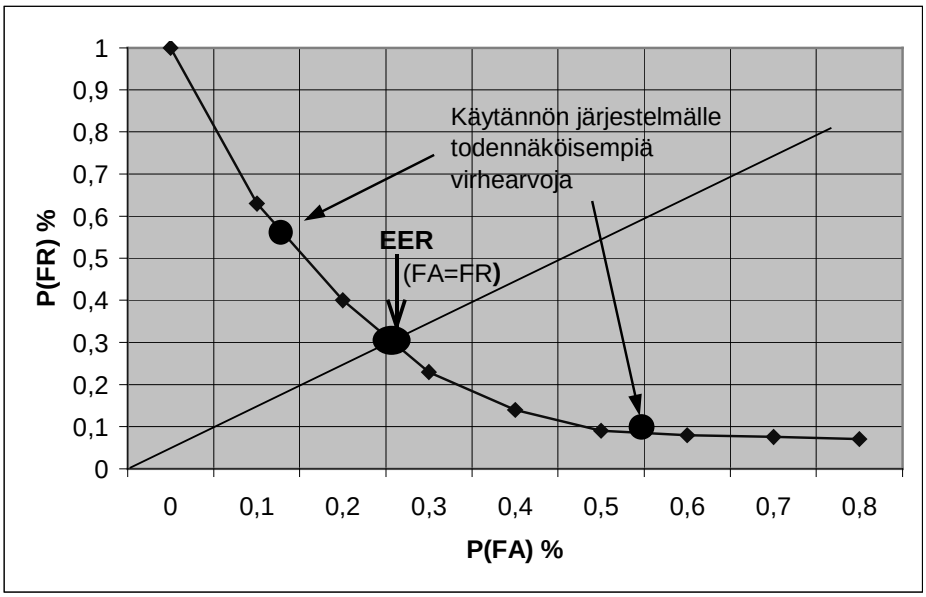
3.3 ROC (Receiver Operating Characteristics) ja EER (Equal Error Rate)

Yhdistämällä virhekäyrien *FA* ja *FR* kuvaajat samaan koordinaatistotasoon ja etsimällä käyrien leikkauskohta, saadaan selville kyseisen puhujanvarmennusjärjestelmän keskimääräisen virhealttius, eli EER [8]. Tällaista kuvaa sanotaan ROC-käyräksi (Kuva 4.).

Keskimääräinen virhealttius saadaan toisin sanoen selville kohdasta, jossa FA = FR. Esimerkiksi järjestelmässä jossa FR-tyyppisten virheiden määräksi on saatu 2.5 prosenttia ja FA-tyyppisten virheiden määräksi samoin 2.5 prosenttia kaikista puhujista, tulee EER:n arvoksi myös 2.5 prosenttia.

EER:n käyttöä järjestelmän yleisen virhealttiuden määrittelyssä voidaan kuitenkin mielestäni hieman kyseenalaistaa. Onhan nimittäin tosiasia, että FA- ja FR-tyyppisten virheiden tulee olla erilainen eri järjestelmissä. Esimerkiksi turvallisuuspalvelussa toimivalla järjestelmällä on aivan oleellista saavuttaa FA:n arvoksi FR:a reilusti pienempi suhteellinen arvo: FR voisi kyseisessä järjestelmässä olla jopa lähellä 10 prosenttia, kun taas FA tulee saada ainakin lähelle 0.1 prosentin arvoa, ehkä jopa alle sen. Teoreettisessa mielessä EER kertoo järjestelmän virhealttiuden erittäin hyvin ja suoraviivaisesti, mutta käytännössä se ei kerro mitään järjestelmän sisäisestä toteutuksesta FA ja FR –arvojen välillä. Lisäksi puhujanvarmennusjärjestelmissä vaaditaan to-

dellisuudessa erilaiset kynnysarvot kaikille puhujille [3, 8], joka: 1. paitsi monimutkaistaa käyttäjien lisäämisprosessia järjestelmään 2. vie samalla myös pohjaa yhden, keskimääräisen varmennusjärjestelmän virhearvon (EER), ilmoittavan arviointimittarin käytön järkevyydeltä. Kynnysarvoa voidaan muuttaa valitsemalla ROC-kuvaajalta jokin muu piste EER:n sijasta. Tämä mielestäni myös kannattaa useissa tapauksissa tehdä.



Kuva 4, ROC (Receiver Operating Characteristics) –kuvaaja.

Seuraavaksi tarkastellaan esimerkin vuoksi taulukoita, joihin on laskettu puhujanvarmennusjärjestelmän läpipäästö- ja hylkäysprosentteja mahdollisessa todellisessa järjestelmässä. Taulukoissa käytetään todennäköisyyslaskennan tulosääntöä [11]:

$$P(A \wedge B) = P(A)P(B) \qquad (3.1)$$

Tulosääntö tarkoittaa, että tapahtumien A ja B yhtäaikaisen toteutumisen todennäköisyys on kyseisten tapahtumien todennäköisyyksien tulo.

1. Järjestelmä, jossa vain yksi yrityskerta puhenäytteen antamiseen, EER = 5%:

		Käyttäjä on todellisuudessa:	
		Aito	Huijari
Järjestelmä vastaa:	Aito	.95	.05
	Huijari	.05	.95

Järjestelmä tunnistaa aidon puhujan 95 prosentin todennäköisyydellä ja väittää aitoa käyttäjää huijariksi 5 prosentin todennäköisyydellä. Lisäksi järjestelmä päästää huijarin sisälle 5 prosentin todennäköisyydellä ja tunnistaa huijarin huijariksi 95 prosentin todennäköisyydellä.

2. Järjestelmä, jossa kolme yrityskertaa puhenäytteen antamiseen, puhujat aitoja puhujia, EER = 5%:

Yrityskerrat:	Todennäköisyydet:	Tulokset:
Läpäisy:	.95	.95
Hylkäys, läpäisy	.05 * .95 =	.047
Hylkäys, hylkäys, läpäisy	.05 * .05 * .95 =	.002
Hylkäys, hylkäys, hylkäys	.05 * .05 * .05 =	.0001

Järjestelmä tunnistaa aidon puhujan ensimmäisellä yrityskerralla 95 prosentin todennäköisyydellä. Toisella kerralla tunnistetuksi tuleminen vaati ensin yhden hylkäyksen, joka tapahtuu 5 prosentin todennäköisyydellä. Näin todennäköisyys tälle tapahtumalle on 0.5 * 0.95 = 0.047, eli 4.7 prosenttia. Vastaavasti kolmen peräkkäisen hylkäyksen todennäköisyys aidolle puhujalle on 0.5 * 0.5 * 0.5 = 0.0001, eli 0.01 prosenttia.

3. Järjestelmä, jossa kolme yrityskertaa puhenäytteen antamiseen, puhujat huijareita, EER = 5%:

Yrityskerrat:	Todennäköisyydet:	Tulokset:
Läpäisy	.05	.05
Hylkäys, läpäisy	.95 * .05 =	.047
Hylkäys, hylkäys, läpäisy	.95 * .95 * .05 =	.045
Hylkäys, hylkäys, hylkäys	.95 * .95 * .95 =	.857

Järjestelmä päästää huijarin ensimmäisellä yrityskerralla sisään 5 prosentin todennäköisyydellä. Toisella kerralla tunnistetuksi tuleminen vaatii ensin yhden hylkäyksen, joka tapahtuu 95 prosentin todennäköisyydellä ja virheellisen tunnistuksen toisella yrityskerralla. Näin todennäköisyys tälle tapahtumalle on 0.5 * 0.95 = 0.047, eli 4.7 prosenttia. Vastaavasti kolmen peräkkäisen hylkäyksen todennäköisyys huijarille on 0.95 * 0.95 * 0.95 = 0.857, eli 85.7 prosenttia.

3.4 VT (Verification Throughput)

Puhujanvarmennusjärjestelmän toimivuutta voidaan mitata myös mittaamalla järjestelmän kokonaissuoritusaikoja todellisen käyttötilanteen mukaisilla syötteillä. Järjestelmän kokonaissuoritusaika koostuu seuraavista suorituksen osa-ajoista [5]:

- Ajasta, joka järjestelmällä kuluu hyväksyttävän ääninäytteen saamiseen. Ts. kuinka useasti käyttäjä joutuu uusimaan ääninäytteensä, eli kuinka monta *FR* -tyyppistä virhettä järjestelmä tekee ennen puhujan onnistunutta varmennusta.

- Ajasta, joka järjestelmällä kuluu hyväksyttävän ääninäytteen varmentamiseen. Tähän osavaiheen kuluva aika riippuu järjestelmän toteutusratkaisusta (onko kyseessä esimerkiksi VQ-algoritmeihin vai neuroverkkoihin pohjautuva tunnistusjärjestelmä), järjestelmän käyttöympäristöstä, sekä siitä kuinka pitkälle järjestelmän toteutus on optimoitu.

- Ajasta, joka järjestelmällä kuluu vastauksen antamiseen käyttäjälle. Vastausaika saattaa olla pitkä etenkin järjestelmissä, jotka vastaavat käyttäjälle puhesyntetisaattorin välityksellä. Aikaa kuluu tällöin varsinaisen puhujan tunnistamisen lisäksi puheen syntetisointiin.

3.5 Puhujan varmennusjärjestelmien vahvuudet ja heikkoudet

Vertailtaessa puhujanvarmennusjärjestelmiä muihin tunnistusjärjestelmiin täytyy arviointeihin ottaa odotettavissa olevien hyötyjen lisäksi mukaan myös järjestelmien mahdolliset heikkoudet.

Vahvuuksia:

- Edullisuus: Useimmat puhujanvarmennusjärjestelmät toimivat pääpiirteissään ohjelmistotasolla. Tästä syystä tunnistusjärjestelmän käyttöä varten ei tarvitse välttämättä hankkia uutta kallista teknologiaa. Automaattiset puhujanvarmennusjärjestelmät ovatkin edullisia verrattuna useimpiin muihin käyttäjän henkilökohtaisia piirteitä identifiointitehtävään käyttäviin järjestelmiin. Mahdollisia pakollisia, mutta suhteellisen edullisia investointeja voivat olla hyvälaatuinen mikrofoni ja digitaaliseen signaalinkäsittelyyn *DSP* (*Digital Signal Processing*) tarvittava piiritason toteutus.
- Turvallisuus: Toisen puhujan ääntä on äärimmäisen vaikea matkia riittävän hyvin mahdollista varmennusjärjestelmän väärinkäyttöä ajatellen [3, 5]. Tämä koskee niin puheäänen yleispiirteitä (murre, puhetyyli, puheen tunneperäiset yksityiskohdat, …), kuin matalammankin tason piirteitä (äänenkorkeus, äänen spektrin suuruusluokka, formanttien taajuudet, …).
- Helppokäyttöisyys: Varmennusjärjestelmän opetusvaiheen jälkeen käyttäjän tarvitsee usein ainoastaan toistaa järjestelmän vaatima lause ilman tarkempaa tietämystä varmennusjärjestelmän syvemmän tason toteutuksesta. Lisäksi monet käyttäjät kokevat puhujanvarmennusjärjestelmät esimerkiksi sormenjälkien- tai silmänrakenteen tunnistamiseen pohjautuviin järjestelmiin verrattuna vähemmän henkilön yksityisyyttä loukkaaviksi.

Heikkouksia:

- Varmennuksen aikavaativuus: Henkilöllisyyden varmentaminen automaattisen puhujanvarmennusjärjestelmän avulla vaatii usein enemmän käyttäjän aikaa kuin esimerkiksi tietoko-

neen näppäimistöltä kirjoitettavat PIN-koodit ja salasanat. Toisaalta jatkuva teknologian kehitys vähentää jatkuvasti tätä tunnistusjärjestelmien välistä aikaeroa.

- Varmennuksen tehokkuus: Puhujanvarmennusjärjestelmien tehokkuuteen vaikuttavat tekijät ovat erittäin herkkiä ulkopuolisille häiriöille, joita ovat esimerkiksi ympäristön melu, käytettävän tiedonsiirtokanavan häiriöt ja puhujanvarmennusjärjestelmään liitetyn mikrofonin huono laatu [5]. Käyttäjän muuttuvat ominaisuudet tuottavat myös ongelmia varmennusjärjestelmille; käyttäjän ääni muuttuu iän myötä pitkällä aikavälillä, mutta vaikeampi tilanne on esimerkiksi flunssasta johtuva äänen muuttuminen.
- Varmistuksen luotettavuus: Toisin kuin salasanoihin tai PIN-koodiin pohjautuvissa henkilönvarmennusjärjestelmissä, puhujanvarmennusjärjestelmillä ei ehkä koskaan päästä tilanteeseen, jossa huijari ei saisi huijattua järjestelmää [3].

4 Päätöslogiikan suunnittelu

Puhujanvarmennusjärjestelmien sisältämä päätöslogiikka on järjestelmien toiminnan kannalta avainasia. Pienetkin puutteet päätöslogiikassa vaikuttavat suoraan järjestelmän luotettavuuteen ja voivat romuttaa sitä kautta koko varmennusjärjestelmän käytettävyyden esimerkiksi turvallisuuspalvelutehtävissä. Puhujanvarmennusjärjestelmien päätöslogiikan suunnitteluun ja kehittämiseen käytetään samoja työkaluja kuin järjestelmien evaluointiinkin. Oikean kynnysarvon määrittely on tärkeä osa järjestelmän päätöslogiikkaa ja varmennuspäätöstä.

4.1 Varmennuspäätös ja kynnysarvon (threshold value) selvittäminen

Varmennuspäätöksen tekeminen perustuu varmennusta haluavan puhujan piirrevektoreiden- ja väitetyn puhujan mallin väliseen vertailuun ja varmennuspäätös tehdään vertailussa todetun samankaltaisuuden pohjalta [8]. Järjestelmän päätöslogiikkaan asetettu kynnysarvo kertoo järjestelmälle, kuinka samankaltaisia puhujamallien tulee olla hyväksytyn varmennustuloksen saamiseen, tai toisaalta kuinka erilaisia puhujamallien tulee olla varmennuksen hylkäämiseen. Kynnysarvon asettaminen on yksi suurimmista ongelmista tosielämän puhujanvarmennusjärjestelmissä [10].

Jos puhujamalli on toteutettu stokastisiin malleihin pohjautuvan päätöslogiikan avulla, varmennustehtävä on itse asiassa todennäköisyyksien vertailua [2]. Puhuja hyväksytään, mikäli aidon puhujan todennäköisyys on suurempi kuin huijarin todennäköisyys.

$$P(K_X) > P(\overline{K_X}) \tag{4.1}$$

Kaavassa K on puhujien joukko ja K_X tietty yksi puhuja joukosta K. Vastaavasti $\overline{K_X}$ tarkoittaa, että puhuja on joku *vastinjoukon* (cohort set) edustaja. Vastinjoukko tarkoittaa tässä sitä joukkoa, johon kuuluvat kaikki muut puhujat kuin tämä nimenomainen varmennettava puhuja [8].

Kynnysarvon merkitystä puhujanvarmennustehtävässä voidaan havainnollistaa seuraavan puhujanvarmennustehtävän *päätössäännön* [8] avulla:

$$W = \begin{cases} hyväksy & ,\log \dfrac{P(K_X)}{P(\overline{K_X})} > \delta_X \\ hylkää & ,muulloin \end{cases} , \qquad (4.2)$$

missä W on varmennustulos ja δ_X on väitetylle puhujalle käytetty kynnysarvo.

Kynnysarvon määrittäminen vaatii ohjelmoijalta tietämystä järjestelmän käytön sovelluskohdealueesta [8]: Kynnysarvon sallittuun *toleranssiin*, mittapoikkeamaan, vaikuttaa nimittäin käytettävän varmennusjärjestelmän yleisen tarkkuuden lisäksi tehtävä, johon järjestelmä on liitetty. Kynnysarvo on yleensä tarkempi, eli vaikeammin läpi päästävä esimerkiksi teollisuuslaitoksen ovenavausjärjestelmässä kuin kotitietokoneeseen liitetyssä sisäänkirjautumisjärjestelmässä [5]. Kynnysarvo on yleensä järjestelmäkohtaisuuden lisäksi myös käyttäjäkohtainen [8]. Tämä tarkoittaa että jokaiselle käyttäjälle joudutaan määrittelemään omakohtainen kynnysarvo järjestelmän opetusvaiheessa.

Kynnysarvon asettaminen on ongelmallista myös sen vuoksi, että arvon määrittäminen tehdään usein pelkästään opetusvaiheessa kerätyn puhedatan (speech corpora) perusteella [7]. Etukäteen määritettyä, opetusdatan pohjalta luotua kynnysarvoa kutsutaan *a priori*-kynnysarvoksi. Kynnysarvon määritteleminen etukäteen määrää samalla pitkälle järjestelmän tehokkuuden [7]. Esimerkiksi EER:n pohjalta määritetty kynnysarvo on *a priori*-kynnysarvo. EER:n pohjautuva kynnysarvon määritys sopii hyvin sisällöstä riippuvaan varmennusjärjestelmään, koska se varmistaa optimaalisen varmennustehokkuuden opetusvaiheessa käytetylle tietojoukolle [7]. Sitä vastoin sisällöstä riippumattomia varmennusjärjestelmiä ajatellen EER:ään pohjautuva kynnysarvon määritys ei usein tuota optimaalista ratkaisua. Tämä johtuu siitä, että EER:äänn pohjautuva kynnysarvo vaatii tehokkaasti toimiakseen saman sisällön omaavat puhenäytteet sekä opetusvaiheessa että tunnistusvaiheessa [7]. Ongelmaksi tässäkin tulee puhujan äänen nopeat muutokset (flunssa, tms.). Optimaalisessa varmennusjärjestelmässä puhujan mallia kehitetään jatkuvasti järjestelmän normaalin käytön yhteydessä [7].

Varmennusjärjestelmien tehokkuutta voidaan parantaa käyttämällä erilaisia opetusdatan optimointimenetelmiä. Menetelmien ideana on minimoida opetusdatan sisältämä, varmennuksessa epäoleellinen informaatio. Menetelmät on tarkoitettu pääasiassa sisällöstä riippumattomiin järjestelmiin ja niiden käyttö parantaa todistetusti kyseisten järjestelmien varmennustarkkuutta [7]. Eräänä menetelmänä on J.H. Liun ja K. Chenin [7] kehittämä menetelmä, jonka perusajatuksena on poistaa opetusdatasta äänityslaitteiston, sekä käyttäjän tunne- ja terveystilanteen aiheutuvat vääristymät. Idea perustuu ajatukselle, että opetusdataa on usein hyvin rajallinen määrä, joten se kuvaa hyvin vain osaa puhujista. Joidenkin käyttäjien osalle voi sattua esimerkiksi väärin kohdistettu mikrofoni tai käyttäjälle flunssainen päivä.

5 Yhteenveto

Automaattinen puhujanvarmennus on jo tällä hetkellä käyttökelpoinen vaihtoehto moneen henkilöllisyydenvarmentamistilanteeseen. Järjestelmien sisältämän päätöslogiikan kehittäminen etenkin kynnysarvojen valinnan kannalta ansaitsee paljon tutkimusta jatkossakin. Automatisoitu kynnysarvojen määrittäminen eri puhujille, järjestelmän keskimääräisen varmennuskyvyn siitä kärsimättä, on yksi varmennusjärjestelmien kehitykseen kohdistuvista suurista haasteista tulevaisuudessa. Puhujakohtaisten kynnysarvojen määrittely on suurempi ongelma sisällöstä riippumattoman-, kuin sisällöstä riippuvan järjestelmän tapauksessa. Syynä tähän on opetusvaiheessa kerätyn datan vähäisyys. Sisällöstä riippuvan järjestelmän kynnysarvojen laskentaan EER tuo nopean ja luotettavan menetelmän.

Puhujanvarmennusjärjestelmien päätöslogiikan automatisointi voisi tulevaisuudessa mahdollistaa kynnysarvon automaattisen muokkaamisen käyttäjän jokaisella sisäänkirjautumiskerralla kerätyn puhedatan perusteella. Näin järjestelmän käytettävyys pysyisi koko ajan korkeana, riippumatta käyttäjän puheäänessä tapahtuvista nopeistakaan muutoksista (käyttäjän vilustuminen, tms.). Käytettävyydellä tarkoitan tässä järjestelmän antamien FR-tyyppisten virheiden minimointia myös turvallisuuskäyttöön tarkoitetussa varmennusjärjestelmässä.

Viiteluettelo

[1] Biometrics, http://www.biometrics.org, Internet-sivu. Viitattu 7.9.2001.

[2] Bourland, H., Morgan, N.: ”Speaker Verification – A Quick Overview”, tutkimusraportti, Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), 1998.

[3] Campbell, P. Joseph: ”Speaker Regocnition: A Tutorial”, *Proceedings of the IEEE*, 85(9), September 1997, pp. 1437 –1462.

[4] Deller Jr. J. R., Proakis J.G., Hansen J.H.L: *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, New York, 1993.

[5] Jayant M. Naik, ”Speaker Verification: A Tutorial”, *IEEE Communications Magazine*, January 1990, pp. 42-48.

[6] Kinnunen, T: *Automaattinen puhujan tunnistus*, Pro gradu –tutkielma, Joensuun yliopisto, Tietojenkäsittelytieteen laitos, 1999.

[7] Liu, J. H., Chen, K.:”Pruning Abnormal Data for Better Making A Decision In Speaker Verification”, *Proceedings of 6th International Conference on Spoken Language Processing (ICSLP’2000)*, pp. 1005-1008, Beijing, China, 2000.

[8] Lötjönen, M.: *Ääneen perustuva käyttäjän todentaminen puhelinverkon lisäarvopalveluissa*, Diplomityö, Lappeenrannan teknillinen korkeakoulu, Tietotekniikan osasto, 2001.

[9] Niemi-Laitinen T.: *Puhujantunnistus rikostutkinnassa*, Yleisen fonetiikan lisensiaatintutkimus, Fonetiikan laitos, Helsingin yliopisto, 1999.

[10] Pierrot, J.-B., Lindberg, J., Koolwaaij, J., Hutter, H.-P., Genoud, D., Blomberg, M., Bimbot, F.: ”A Comparison Of a Priori Threshold Setting Procedures For Speaker Verification In The CAVE Project”, *Proceedings of Acoustics, Speech and Signal Processing (ICASSP’1998)*, pp. 125 – 128, Seattle, USA, 1998.

[11] Scheaffer, R.L.: *Introduction to Probability and Its Applicatications*, Duxbury Press,1990.