(19) SUOMI - FINLAND
(FI)

PATENTTI- JA REKISTERIHALLITUS
PATENT- OCH REGISTERSTYRELSEN
FINNISH PATENT AND REGISTRATION OFFICE

(10) **FI 954527 A7**

(12) **JULKISEKSI TULLUT PATENTTIHAKEMUS**
PATENTANSÖKAN SOM BLIVIT OFFENTLIG
PATENT APPLICATION MADE AVAILABLE TO THE PUBLIC

(21) Patenttihakemus - Patentansökan - Patent application 954527

(51) Kansainvälinen patenttiluokitus - Internationell patentklassifikation - International patent classification (IPC[7])
G10L 15/06

(22) Tekemispäivä - Ingivningsdag - Filing date **25.03.1994**

(23) Saapumispäivä - Ankomstdag - Reception date **25.09.1995**

(41) Tullut julkiseksi - Blivit offentlig - Available to the public **20.11.1995**

(43) Julkaisupäivä - Publiceringsdag - Publication date **13.06.2019**

(86) Kansainvälinen hakemus - Internationell ansökan - International application **25.03.1994 PCT/GB1994/000629**

(32) (33) (31) Etuoikeus - Prioritet - Priority

25.03.1993 EP 93302302

(71) Hakija - Sökande - Applicant

1 • **British Telecommunications public limited company,** 81 Newgate Street, London EC1A 7AJ, ISO-BRITANNIA, (GB)

(72) Keksijä - Uppfinnare - Inventor

1 • **Pawlewski, Mark,** United Kingdom, ISO-BRITANNIA, (GB)
2 • **Tang, Joseph Gordon,** United Kingdom, ISO-BRITANNIA, (GB)

(74) Asiamies - Ombud - Agent

Leitzinger Oy, High Tech Center, Tammasaarenkatu 1, 00180 Helsinki

(54) Keksinnön nimitys - Uppfinningens benämning - Title of the invention
Menetelmä ja laite puhujan tunnistamiseksi
Förfarande och anordning för identifiering av talare

Menetelmä ja laite puhujan tunnistamiseksi.-
Förfarande och anordning för igenkänning av en talare

Esillä oleva keksintö liittyy puheenkäsittelyyn ja erityisesti prosessointiin puhujan tunnistamiseksi. Tunnistusprosessi käsittää puhujan tunnistuksen, jossa ilmaistaan tai vahvistetaan puhujan henkilöllisyys, ja puheen tunnistuksen, jossa ilmaistaan määrätty sana (tai joskus, lause tai foneemi, tai muuta puhuttua asiaa). Puheen tunnistus sisältää niin sanotun puhujasta riippumattoman tunnistuksen, jossa tunnistusprosessoinnissa käytetään useammalta puhujalta saatua puhedataa. Puheen tunnistuksessa käsittelyssä pyritään yleensä vähentämään eri puhujien lausumien sanojen vaikutusta, kun taas puhujan tunnistuksessa asia on päinvastoin.

Tunnistusprosessoinnissa on tavallista, että puhedata, tyypillisesti digitaalisessa muodossa, syötetään niin sanottuun edustaprosessoriin, joka tulevan puhedatan virrasta johtaa tiivistetymmän, havainnoinnin kannalta merkittävän datajoukon, jota sanotaan edustaominaisuusjoukoksi tai -vektoriksi. Puhetta syötetään esimerkiksi tyypillisesti mikrofonin kautta, siitä otetaan näytteitä jotka digitoidaan ja jaetaan kehyksiksi, joiden pituus on 10 - 20 ms (näytteenotto esimerkiksi 8 kHz:llä), ja jokaisen kehyksen osalta lasketaan K kertoimen joukko (jolloin kertoimia on tyypillisesti 5 - 25). Koska sanaa kohti on N kehystä, esimerkiksi 25 - 100, niin ominaisuusvektorissa on N x K (suuruusluokkaa 1000) kerrointa. Puhujan tunnistuksessa tunnistettavan puhujan oletetaan yleensä lausuvan ennalta määrätyn sanan, jonka tunnistuslaite ja puhuja tuntee (esimerkiksi PIN-numero pankkialalla). Sanan talletettu esitys, joka tunnetaan mallina, käsittää tämän sanan vertailu-ominaisuusmatriisin, joka on aikaisemmin johdettu oikeaksi tunnetun puhujan avulla. Tunnistettavan puhujan osalta johdettua ja syötettyä ominaisuusmatriisia verrataan malliin, ja näiden kahden välistä yhtäläisyyden mittaa verrataan kynnysarvoon hyväksymispäätöstä varten.

Eräs ongelma syntyy siitä, että puhujilla on taipumus muuttaa sanojen ääntämisnopeutta, niin että määrättyä sanaa vastaava syötetty puhematriisi voi olla pidempi (käsittää esimerkiksi useampia kehyksiä) tai lyhyempi kuin tämän sanan malli. Tämän vuoksi tunnistuslaitteen on aika ajoin tehtävä näiden kahden matriisin ajallinen kohdistus, ennenkuin vertailu voidaan tehdä, ja eräs tällainen varsin tunnettu ajallisen kohdistamisen ja vertailun menetelmä on Dynamic Time Warp (DTW) -menetelmä, jota on selitetty esimerkiksi artikkelissa Rabiner, ym: Speaker Independent Recognition of words using Clustering Techniques; IEEE Trans. on ASSP, vol 24, no. 4, elokuu 1979.

Tunnistusprosessia varten on käytetty tai ehdotettu erilaisia ominaisuuksia. Koska puheentunnistukseen käytetyt ominaisuuden on tarkoitettu erottamaan sanat toisistaan olematta puhujalle herkkä, kun taas puhujan tunnistukseen käytetyt ominaisuudet on tarkoitettu erottamaan puhujat toisistaan tunnetun sanan tai tunnettujen sanojen osalta, niin yleensä toisentyyppiseen tunnistukseen sopiva ominaisuus on sopimaton toiseen tunnistukseen. Eräitä ominaisuuksia puhujan tunnistusta varten on selitetty artikkelissa Atal: Automatic Recognition of Speakers from their voices; Proc IEEE vol 64, s. 460 - 475, huhtikuu 1976.

Tunnettujen kertoimien eräs tyyppi on cepstrum. Cepstrumit muodostetaan suorittamalla spektrin hajautus (esimerkiksi spektrimuunnos, kuten Fourier-muunnos) ottaen muunnoskertoimien logaritmi, ja suorittamalla käänteinen spektrin kokoaminen. Puhujan tunnistuksessa tunnetaan LPC-cepstrum (Linear Prediction Coefficient) ja FFT-cepstrum (Fast Fourier Transform), joista edellinen on yleisemmin käytetty.

Puheen tunnistuksessa MFCC (Mel-Frequency Cepstrum Coefficient) on tunnettu ominaisuus. Algoritmi MFCC-kertoimien laskemiseksi ja etäisyysmitan laskemiseksi MFCC-ominaisuuden ja sanan mallin välillä DTW-menetelmää käyttäen on esitetty artikkelissa Chollet ja Gagnoulet: On the evaluation of Speech Recognisers and Data Bases using a Reference System; 1982 IEEE International Conference on Acoustics, Speech and Signal Processing, s. 22026 - 2029, joka liitetään tähän kokonaisuudessaan (siinä olevat viittaukset mukaanlukien).

MFCC-ominaisuusvektori johdetaan yleensä suorittamalla spektrimuunnos (esim. FFT) puhesignaalin jokaista kehystä varten niin että saadaan signaalispektri; spektrin termit integroidaan sarjaksi laajoja kaistoja, jotka jakaantuvat taajuusakselia pitkin epätasaiseksi, niin sanotuksi 'mel-taajuusjaoksi'; jokaisessa kaistassa otetaan suuruuden logaritmit; ja sitten suoritetaan toinen muunnos (esim. DCT, Discrete Cosine Transform - muunnos) MFCC-kerroinjoukon muodostamiseksi kehystä varten. On havaittu, että käyttökelpoinen tieto rajoittuu yleensä alemman kertaluokan kertoimiin. Mel-taajuusjako voi esimerkiksi muodostua taajuuskaistoista, jotka jakaantuvat tasaisesti lineaariselle taajuusasteikolle välillä 0 - 1 kHz, ja tasaisesti logaritmiselle taajuusasteikolle yli 1 kHz taajuuksilla.

MFCC-kertoimet poistavat tiedon äänenkorkeudesta, joka on käyttökelpoinen puheen tunnistamiseksi, koska se vaihtelee eri puhujilla, mutta sitä ei toivota puhujan tunnistamiseksi. Vastaavasti MFCC-kertoimia ei ole pidetty edullisina puhujan tunnistuksessa.

Artikkelissa S. Furui: Speaker Recognition by Statistical Features and Dynamic

Features; Review of Electrical Communications Laboratories, Vol 30, No 3, 1982, s. 467 - 482 kuvataan ja verrataan toisiinsa neljää puhujan tunnistusmenetelmää. Kahdessa näistä menetelmistä spektrivaippakäyrä on edustettu logaritmisilla pinta-alaosuuksilla, ja kahdessa toisessa LPC-cepstrum-kertoimien avulla. LPC-cepstrum-kertoimien käytön tapauksessa muodostetaan kertoimien keskiarvo koko lauseen ajalta, joka voi käsittää useita sanoja (esimerkiksi jopa 7 sanaa), ja keskiarvoistetut arvot vähennetään cepstrum-kertoimista jokaisessa kehyksessä, niin että kompensoidaan taajuusvastevääristymät, joita siirtojärjestelmä aiheuttaa. Sen jälkeen normalisoitujen cepstrum-kertoimien aikafunktiot laajennetaan ortogonaalisen polynomiesityksen avulla 90 ms:n jaksolle aina 10 ms:n jaksoissa. Polynomien käsittely tuottaa polynomiesityksen, joka vastaa jokaisen segmentin jokaisen cepstrum-kertoimen keskivertoa, kaltevuutta ja kaarevuutta. Cepstrum-kertoimien alkuperäisiä aikafunktioita käytetään nollannen kertaluokan polynomikertoimien sijasta, yhdessä ensimmäisen ja toisen kertaluokan polynomikertoimien kanssa, niin että jokaista lausetta edustaa 30-ulotteisen vektorin aikafunktio. Näistä 30 elementistä valitaan 18 elementin joukko tarkoituksella laajentaa asiakkaan ja huijarin näytelausahdusten (jotka on määritetty pidemmän opetus/rekisteröintivaiheen kuluessa) välistä etäisyysjakautumaa. Vaihtoehtoisessa käsittelyjärjestelyssä LPC-analyysiä seuraa tilastollinen ominaisuuksien erotusprosessi, joka käsittää mm. Fourier:n kosini-laajennuskertoimien muodostamisen. Tällä vaihtoehtoisella käsittelyllä lopullinen ominaisuusjoukko käsittää 60 valittua elementtiä. Tällä viimeksimainitulla käsittelyjärjestelyllä ("Method 2"), jossa käytettiin tilastollisia ominaisuuksia mukaanlukien Fourier-kertoimia, jotka otettiin LPC-cepstrum-kertoimien aikafunktioista ja perustaajuudesta, ilmoitettiin tuotettavan tunnistustarkkuus 99.9 kyseessä olevalla opetusdatalla ja huijarin datalla. Furui päättelee, että tulokset "osoittavat, että LPC-cepstrum-kertoimet ovat paljon tehokkaampia kuin logaritmiset pinta-alasuhteet". Furui ei esitä mitään opetusta muiden cepstrum-kertoimien käyttämiseksi, eikä tulevan puheen suodattamista useammalle ennalta määrätylle taajuuskaistalle. Lisäksi Furui epäsuorasti opettaa, että on arvokasta käyttää hyvin suurta ominaisuusjoukkoa (esim. jopa 60 elementtiä) - jonka välttäminen on selvästikin toivottavaa, erityisesti kun kyseeseen tuleva tunnistajan populaatio on hyvin suuri.

Artikkelissa J.M. Naik ym: Speaker Verification over Long Distance Telepohone Lines; ICASSP 89, Vol 1, 23.5.1989, s. 524 - 527, vertaillaan puhujan hyväksymismenetelmiä, joissa käytetään joko malliin perustuvaa DTW-prosessia (Dynamic Time Warping) tai HMM-prosessia (Hidden Markov Modelling). Tässäkin suoritettiin LPC-analyysi, jolla saatiin esikäsitelty tieto, josta vuorostaan ominaisuudet johdettiin. Aluksi laskettiin LPC-datan jokaisesta kehyksestä 32 parametria:

- puheen tason estimaatti, dB,

- kehyksen tehollinen energia, dB (rms),
- spektrimuutoksen nopeuden skalaarimitta,
- suodatinryhmän voimakkuutta, dB,
  - mel-etäisyyksin simuloidut suodatinryhmät,
  - kehyksen energialla normalisoidut ryhmät,
- kehysten energian aikaerotus 40 ms:n ajalta,
- suodatinryhmän voimakkuuden aikaerotus 40 ms:n ajalta.

Puheen ominaisuudet, joita käytettiin verrattaessa referenssi- ja koemalleja, saatiin näiden 32 parametrin lineaarisen muunnoksen kautta, niin että muodostettiin 18 ominaisuutta mallina käytettäväksi. Mitään selitystä tai syytä ei anneta mel-etäisyyksin olevien suodatinryhmien käytölle, eikä artikkelissa ole mitään ehdotusta LPC- cepstrum-kertoimista poikkeavien muiden kertoimien käyttämiseksi.

Keksinnön eräänä tavoitteena on aikaansaada puhujan tunnistamista varten menetelmä ja laite, joissa käytetään parannettua etupään ominaisuutta.

Vastaavasti keksinnöllä aikaansaadaan puhujan tunnistamista varten menetelmä ja laite, joissa puhesignaalia käsitellään tunnistusominaisuuksien johtamiseksi, jotka käsittävät joukon spektrin amplitudikertoimia jakaantuneena epätasaiselle taajuusasteikolle, ja joista ainakin yksi amplitudikerroin normalisoidaan sen määrätyn aikajakson yli otetun keskimääräisen tason avulla.

Normalisoinnilla pienennetään puhesignaalin mahdollisen pitkäaikaisen suodatuksen vaikutukset, joita voi syntyä esimerkiksi tietoliikennekanavassa, jonka kautta signaali kulkee.

Kertoimet voidaan jakaa esimerkiksi mel-taajuusasteikolle. Tässä tapauksessa ne voivat olla Mel-taajuus -cepstralkertoimia. Normalisointi tehdään tarkoituksenmukaisesti muodostamalla jokaisen kertoimen pitkän aikavälin aritmeettinen keskiarvo ja vähentämällä tämä jokaisen kertoimen arvosta. Koska kertoimet ovat logaritmisia, vähentäminen vastaa jakamista lähdesignaalin geometrisella keskiverrolla jokaisella mel-taajuuskaistalla.

On havaittu, että kun mel-taajuiset cepstral-kertoimet tällä tavalla on normalisoitu, ne tarjoavat luotettavan ominaisuusjoukon puhujan tunnistamiseksi käyttäen vain pientä kertoimien lukumäärää, ja lisäksi että niiden käyttö soveltuu erityisesti tietoliikennesovellutuksiin, koska riippuvuus puhujan ja tunnistuslaitteen välisen tietoliikennekanavan ominaisuuksista pienenee merkittävästi.

Keksinnön muut näkökohdat ja edullisina pidetyt suoritusmuodot ovat tässä ja patenttivaatimuksissa esitetyn mukaiset, jolloin niiden edut ilmenevät seuraavasta

selityksestä.

Keksintöä selitetään seuraavassa pelkän esimerkin muodossa oheiseen selityksiin ja piirustuksiin viitaten, joissa:

kuva 1 esittää kaaviollisesti keksinnön mukaisen tunnistusprosessorin soveltamisen tietoliikenneympäristössä;

kuvassa 2 on lohkokaavio, joka esittää kaaviollisesti keksinnön erään suoritusmuodon mukaisen tunnistusprosessorin toiminnalliset elementit;

kuvassa 3 on vuokaavio, joka kaaviollisesti esittää kuvan 2 osan muodostavan MFCC-generaattorin toimintaa;

kuvassa 4 on havainnollistava käyrästö taajuusalueelta, jolla havainnollistetaan osaa kuvan 3 prosessista;

kuvassa 5 on vuokaavio, joka esittää yksityiskohtaisemmin äärikohdan ilmaisimen, joka muodostaa osan kuvasta 2;

kuvassa 6 on havainnollistava amplitudikäyrästö ajan funktiona, jolla havainnollistetaan kuvan 5 prosessia;

kuvassa 7 on vuokaavio, joka esittää kaaviollisesti normalisointiprosessorin toiminnan, prosessorin muodostaessa edullisessa suoritusmuodossa osan kuvasta 2;

kuvassa 8a on vuokaavio, joka havainnollistaa kuvan 2 osan muodostavan vertailuprosessorin toimintaa puhujan vahvistamisen suoritusmuodossa;

kuvassa 8b on vuokaavio, joka havainnollistaa kuvan 2 osan muodostavan vertailuoprosessorin toimintaa puhujan tunnistamisen suoritusmuodossa;

kuvassa 9 on vuokaavio, joka esittää kuvan 2 osan muodostavan normalisointiprosessorin toiminnan suoritusmuodossa, joka on vaihtoehtona kuvalle 7;

kuvan 10a käyrästö havainnollistaa MFCC-kertoimien arvoja ajan funktiona kahdella erilaisella tietoliikennekanavalla; ja

kuvan 10b vastaava käyrästö esittää kertoimia, jotka on normalisoitu kuvan 7 edullisena pidetyn suoritusmuodon mukaisesti.

Kuvaan 1 viitaten puhujan tunnistuksen sisältämä tietoliikennejärjestelmä käsittää yleisesti ottaen mikrofonin 1, joka tyypillisesti muodostaa osan puhelimen luurista, tietoliikenneverkon (tyypillisesti yleinen kytkettävä puhelinverkko, PSTN) 2, tun-

nistusprosessorin 3, joka on kytketty vastaanottamaan puhesignaalin verkosta 2, ja tunnistusprosessoriin 3 kytketyn käyttölaitteen 4, joka on järjestetty vastaanottamaan prosessorilta puheen tunnistussignaalin, signaalin osoittaessa tunnistuksen tai muulla tavalla määrätyn puhujan ja ryhtyen toimenpiteisiin sen perusteella. Käyttölaite 4 voi esimerkiksi olla kauko-ohjattu pankkipääte, jolla suoritetaan pankkitoimia.

Monissa tapauksissa käyttölaite 4 muodostaa puhujalle kuultavan vasteen, joka välitetään verkon 2 kautta kaiuttimelle 5, joka tyypillisesti muodostaa tilaajan puhelinluurin osan.

Käytössä puhuja puhuu mikrofoniin 1 ja analoginen puhesignaali välitetään mikrofonista 1 verkkoon 2 ja punnitusprosessorille 3, jossa puhesignaalia analysoidaan ja tunnistuksen osoittava tai muutoin määrätyn puhujan osoittava signaali muodostetaan ja välitetään käyttölaitteelle 4, joka sitten ryhtyy sopivaan toimintaan kun puhuja on tunnistettu.

Tyypillisesti tunnistusprosessorin on saatava dataa, joka koskee puhujan henkilöllisyyttä, ja johon puhesignaalia verrataan, ja tämä datan saaminen voidaan tehdä tunnistusprosessorilla toisessa käyttömuodossa, jossa tunnistusprosessori 3 ei ole kytkettynä käyttölaitteeseen 4, vaan jossa se vastaanottaa puhesignaalin mikrofonilta ja muodostaa tunnistusdatan tätä puhujaa varten. Puhujan tunnistusdatan saamiseksi kuitenkin myös muut menetelmät ovat mahdollisia; puhujan tunnistusdataa voidaan esimerkiksi säilyttää puhujan mukana olevalla kortilla, joka asetetaan kortinlukijaan, josta sitten data luetaan ja välitetään puhelinverkon PSTN kautta tunnistusprosessorille ennen puhesignaalin lähettämistä.

Tyypillisesti tunnistusprosessori 3 ei tiedä tietä, jonka signaali mikrofonilta 1 kulkee verkon 2 läpi; mikrofoni 1 voi esimerkiksi olla kytkettynä matkaviestinjärjestelmän analogisen tai digitaalisen radioyhteyden kautta verkkoon 2, tai se voi tulla toisesta maasta tai monesta eri tyyppisestä ja eri laatuisesta puhelinluurista. Samaten verkossa 2 voidaan valita jokin siirtotie monen erilaisen joukosta, johon voi kuulua radiolinkit, analogiset ja digitaaliset reitit, jne. Vastaavasti tunnistusprosessoriin 3 saapuva puhesignaali Y vastaa mikrofonilla 1 vastaanotettua puhesignaalia S, johon nyt liittyy mikrofonin 1, verkkoon 2 vievän yhteyden, verkon 2 läpi kulkevan kanavan, ja tunnistusprosessorille 3 johtavan yhteyden siirto-ominaisuudet, jotka voidaan yhdistää ja merkitä yhdeksi siirto-ominaisuudeksi H.

Tunnistusprosessori 3

Kuvassa 2 on esitetty edullisen suoritusmuodon tunnistimen toiminnalliset ele-

mentit. Korkeita korostava suodatin 210 vastaanottaa esimerkiksi näytteenottotaajuudella 8 kHz digitoidun puhe-aaltomuodon 8-bittisten numeroiden jonona ja suorittaa korkeita korostavat suodatusprosessin (suorittamalla esimerkiksi suodatuksen $1 - 0{,}95 \bullet z^{-1}$), niin että korkeampien taajuuksien amplitudi kasvaa. Puhekehysgeneraattori 220 vastaanottaa suodatetun signaalin ja muodostaa peräkkäisten näytteiden kehysjonon. Jokainen kehys voi esimerkiksi käsittää 256 peräkkäistä näytettä, ja jokainen kehys voi limittyä 50 % edeltävän ja sitä seuraavan kehyksen päälle, niin että kehysten pituudeksi tulee 32 ms, ja ne esiintyvät taajuudella 1 kehys/16 ms. Voidaan esimerkiksi täyttää rinnakkain 16 ms:n viive-erotuksella kahta kehyspuskuria 221, 222, jotka luetaan vuorotellen.

Jokaisen kehyksen alussa ja lopussa olevien epäjatkuvuuksien aiheuttamien satunnaisten taajuushäiriöiden välttämiseksi jokainen kehys johdetaan sitten Hamming-ikkunprosessorin 223 kautta, joka (sinänsä tunnetulla tavalla) skaalaa näytteet alaspäin kohti kulloisenkin ikkunan reunaa.

Jokaista 256 näytteen kehystä käsitellään sitten MFCC-generaattorilla 230, joka erottaa joukon MFCC-kertoimia (esimerkiksi 8 kerrointa). Samalla jokainen ikkunoitu kehys johdetaan äärikohdan ilmaisimeen 240, joka ilmaisee puhutun lauseen alun ja lopun ja syöttää puhe/ei-puhe -ohjaussignaalin normalisoijalle 250, joka käsittää kerroinmuistin 251 ja normalisointiprosessorin 252. Vastaanotettuaan 'puheen alku'-signaalin äärikohdan osoittajalta 240, normalisoija 250 tallettaa jokaisen peräkkäisen kehyksen osalta joitakin kertoimia tai kaikki 8 kerrointa normalisoinnin kerroinmuistiin 251, kunnes äärikohdan osoittajalta 240 vastaanotetaan 'puheen loppu'-signaali. Tällöin normalisointiprosessori 252 laskee jokaisen puhekehyksen osalta muistiin 251 talletettujen kertoimien perusteella aritmeettisen keskiverto-kerroinarvon jokaista 8 kerrointa varten. Sen jälkeen jokaisen kehyksen osalta jokaisen kertoimen aritmeettinen keskiverto-kerroinarvo vähennetään kulloisestakin talletetusta kerroinarvosta, niin että saadaan normalisoitu matriisi, joka käsittää 8 x N kerrointa (jossa N on kehysten lukumäärä puhutun lauseen alku- ja loppukohdan välillä).

Tämä normalisoitu kerroinmatriisi johdetaan vertailuprosessorille 260, joka lukee annettuun puhujaan liittyvän vastaavan matriisin puhujan mallista 270, suorittaa näiden kahden matriisin vertailun, ja muodostaa tunnistus/ei-tunnistus -lähtösignaalin normalisoidun puhevektorin ja puhujan mallimuistista 270 luetun puhujan mallin välisestä yhteneväisyydestä riippuen.

Korkeiden korostussuodatin 210, ikkunaprosessori 223, MFCC-generaattori 230, äärikohdan osoittaja 240, normalisointiprosessori 252 ja vertailuprosessori 260 voidaan toteuttaa yhdellä tai useammalla digitaalisella signaaliprosessorilaitteella

(DSP, Digital Signal Processor) ja/tai mikroprosessoreilla, jotka on sopivasti ohjelmoitu, jolloin kehyspuskurit 221, 222, kerroinmuisti 251 ja puhujan mallimuisti 270 on järjestetty niihin liittyviin luku/kirjoitus-muistilaitteisiin.

MFCC-kertoimien muodostaminen

Kuvaan 3 viitaten MFCC-generaattorin 230 suorittama prosessi käsittää Fourier-muunnoksen suorittamisen jokaiselle kehykselle; 256 muunnoskertoimen muodostamisen vaiheessa 401; puhesignaalin tehospektrin muodostamisen Fourier-kertoimien avulla summaamalla jokaisella taajuudella reaali- ja imaginaarikomponenttien neliöt vaiheessa 402; tehospektrin integroimisen 19 taajuuskaistan yli vaiheessa 403, kuten alla tarkemmin selitetään kuvaan 4 viitaten, niin että saadaan 19 kaistan tehokerrointa; logaritmin muodostamisen (kantalukuna esimerkiksi 10) jokaiselle 19 kertoimelle vaiheessa 404; diskreetti kosinimuunnoksen suorittamisen 19 logaritmiarvolle vaiheessa 405; ja 8 alimman kertaluokan kertoimen valitsemisen vaiheessa 406.

Diskreetti kosinimuunnos on sinällään tunnettu, ja sitä on esimerkiksi selitetty edellä mainitussa Chollet:n ja Gagnoulet:n artikkelissa; lyhyesti sanoen M_m:n N:s kosinikomponentti saadaan kaavasta:

$$Y_m = \frac{\sum_{n=1}^{N} A(n) \cdot Cos(m \prod (n + 0{,}5))}{N}$$

jossa N on diskreettien taajuuskaistojen lukumäärä (tässä tapauksessa 20, jolloin taajuusalueen kiertoa sovelletaan 20:nnen pisteen saamiseksi) ja A(n) on amplitudi m:nnessä taajuuskaistassa. DCT vaikuttaa dekorreloiden kertoimet A(n) suuressa määrin.

Kuvaan 4 viitaten kuva 4a osoittaa pelkistetysti osan tehospektristä, joka muodostettiin vaiheessa 402. Kuva 4b esittää vastaavan osan mel-taajuuden kolmion muotoisista integrointi-ikkunoista taajuusakselilla. Kolmion muotoiset ikkunat käsittävät välillä 0 - 1 kHz kymmenen ikkunaa, jotka jakaantuvat lineaarisesti taajuusakselille, jolloin jokainen limittyy 50 % viereisen ikkunan päälle, ja yli 1 kHz:llä toiset kymmenen ikkunaa, jotka ovat kolmion muotoiset ja limittyvät 50 % logaritmisella asteikolla.

Kuva 4c esittää kaaviollisesti vaikutuksen, joka saadaan kun pisteittäin kerrotaan tehospektrin jokainen näyte vastaavalla termillä kolmion muotoisessa ikkunassa. Selkeyden vuoksi tässä on esitetty vain ikkunat, joilla on parillinen järjestysnumero.

Seuraavaksi jokaisen ikkunan yli integroidaan kuvan 4c arvot, niin että saadaan tätä ikkunaa vastaava summattu kerroin, kuten kuvassa 4d on esitetty.

Tällä tavalla muodostetut 19 kerrointa (jolloin nollataajuuden kerrointa M_0 ei oteta lukuun) vastaavat täten tehoa, joka voitaisiin kehittää jokaisen kaistanpäästösuodattimen lähtöön vastaavassa kaistanpäästösuodatinten joukossa, jolloin alle 1 kHz suodattimilla on tasaisesti ja yhtä suurin välein jakaantuneet kaistanleveydet, ja yli 1 kHz suodattimilla on tasaisesti ja yhtä suurin välein jakaantuneet kaistanleveydet logaritmisella taajuusasteikolla.

Äärikohdan osoitus

Kuviin 5 ja 6 viitaten seuraavassa selitetään yksityiskohtaisemmin kuvassa 2 olevan äärikohdan osoittajan 240 toimintaa.

Äärikohdan osoittaja 240 muodostaa aluksi jokaisen kehyksen osalta signaalien neliöt ja summat, niin että saadaan kehyksessä olevan tehon tai energian mitta P.

P:n arvoa verrataan ensimmäiseen kynnysarvoon P_L, joka on asetettu suhteellisen alhaiselle tasolle, niin että kohinasignaalit ajoittain voivat ajoittain ylittää sen. Mihinkään toimenpiteeseen ei ryhdytä ennenkuin kehyksen tehotaso on tämän matalan kynnyksen P_L yli. Kun kehys nousee matalan kynnyksen P_L yli, talletetaan lippu, joka osoittaa tälle kehykselle annetun numeron (esitetty kuvassa 5 muuttujana "START").

Kun kehyksen tehon arvo P nousee suuremmaksi kuin yläkynnysarvo P_H, joka vastaa puheen esiintymistä, ja jonka taso on todennäköisen kohinatason yläpuolella, oletetaan että puhetta esiintyy. Puheen alkukohdaksi otettu kohta on kehys, joka esiintyy ennalta määrätyn kehysmäärän ("LEAD") verran ennen sitä kehystä ("START"), jossa signaali nousi alakynnyksen P_L yläpuolelle. Vaikka vahvistetaan että puhetta esiintyy vain signaalitason noustessa yläkynnyksen yläpuolelle, niin tällä tavalla ei menetetä lauseen alkua. Vastaavasti äärikohdan osoittaja 240 tuottaa lähtöönsä tällä tavalla alkupisteeksi lasketun kehyksen numeron normalisoijan 250 ohjausta varten.

Jos puhesignaalin taso jää näiden kahden kynnysarvon välille pidemmäksi ajaksi kuin ennalta määrätty aika T_{max}, niin arvo "START" nollataan.

Kun kehyksen energia laskee yläkynnysarvosta P_H alakynnysarvon P_L alapuolelle, äärikohdan osoittaja 240 odottaa ennalta määrätyn kehysmäärän verran, jota sanotaan "ylitys"-ajaksi T_{oh}. Jos taso jälleen ylitysajan aikana nousee alakynnyksen P_L yli, niin puheen oletetaan esiintyvän edelleen. Kun signaalin tehotaso on laskenut

alakynnyksen P_L alle pidemmäksi kuin T_{oh} kehyksen ajaksi, niin lause oletetaan loppuneeksi, ja äärikohdan osoittaja tuottaa lähtöönsä äärikohdan kehysnumeron, joka vastaa sen hetkistä kehysnumeroa, josta on vähennetty kehysten T_{oh} määrä (eli kohta, jossa signaali viimeksi oli kynnysarvossa P_L) ja johon on lisätty ennalta määrätty lukumäärä kehyksiä, jota sanotaan viiveeksi "LAG".

Normalisointi

Kuvaan 7 viitaten seuraavassa selitetään normalisoijan 250 suorittamaa normalisointiprosessia.

Kehykset, joissa on 8 MFCC-kerrointa kehystä kohti, talletetaan peräkkäin kerroinmuistiin 251. Kun äärikohdan ilmaisija 240 ilmaisee puhutun lauseen loppukohdan, se ilmoittaa alku- ja loppukohdan kehysten numerot normalisoijalle 250. Sitten normalisointiprosessori 252 hakee jokaisen 8 kertoimen osalta muistista vastaavat kertoimen arvot alku- ja loppukohdan välisten kehysten välissä, ja muodostaa aritmeettisen keskivertoarvon summaamalla kerroinarvot ja jakamalla arvolla N, jolloin N on kehysten lukumäärä alku- ja loppukehyksen välissä. Tästä saadaan joukko, jossa on 8 keskimääräistä arvoa $\overline{M_i}$ (jossa i = 1 ... 8).

Seuraavaksi normalisointiprosessori 252 laskee jokaisen kehyksen jokaisen kertoimen osalta normalisoidun kerroinarvon $G_{i,K}$ (jossa K on kehyksen numero) vähentämällä vastaavan keskimääräisen arvon $\overline{M_i}$ jokaisesta kerroinarvosta $M_{i,K}$.

Sen jälkeen normalisointiprosessori 252 tuottaa lähtöönsä 8 x N kerrointa sisältävän joukon, joka muodostaa normalisoidun vektorin $G_{i,K}$.

Vertailuprosessi

Vertailuprosessorin 260 yksityiskohtainen selitys on tarpeeton, koska se toimii tavanomaisella tavalla. Kuva 8a esittää kaaviollisesti vertailuprosessorin toiminnan puhujan tunnistusta varten. Vertailuprosessori lukee ominaisuusvektorin G, joka käsittää normalisoidut MFCC-kertoimet; lukee puhujan mallin T, joka käsittää vastaavan kertoimien referenssivektorin; suorittaa näiden kahden vektorin vertailun käyttäen esimerkiksi sinänsä tunnettua DTW-algoritmia, jolla vektorit kohdistetaan ajallisesti (esimerkiksi käyttäen edellä mainitussa Chollet:n ja Gagnoulet:n artikkelissa esitettyä algoritmia) skalaarisen etäisyysmitan D muodostamiseksi, joka osoittaa näiden kahden vektorin välisen erotuksen; ja vertaa etäisyysmittaa D kynnysarvoon. Jos etäisyys D on pienempi kuin kynnysarvo, puhuja hyväksytään talletettua mallia vastaavana; muussa tapauksessa puhuja hylätään. Kuva 8b esittää vertailuprosessorin 260 vastaavan toiminnan puhujan tunnistamisessa. Tässä tapa-

uksessa mallivarastosta 270 luetaan peräkkäin joukko erilaisia vektoreita T_i, joihin puhevektoria G vuorotellen verrataan vastaavan etäisyysmetricin D_i muodostamiseksi. Sen jälkeen puhuja tunnistetaan vastaten sitä mallia, josta puhevektori vähiten poikkeaa (eli joka tuottaa pienimmän metricin D_i).

Vaihtoehtoiset suoritusmuodot

Edellä esitetyssä suoritusmuodossa, jota tarkasteltiin kuvaan 7 liittyen, normalisoijan 250 on tiedettävä sekä lauseen alku- että loppukohta, ennenkuin se voi laskea arvon N (alku- ja loppukohdan välisten kehysten lukumäärän) ja kerroinarvojen summan M_{TOT}, ja siitä jokaisen kertoimen keskiarvon, ja edelleen jokaisen kertoimen normalisoidun arvon. Vastaavasti normalisoijan 250 on odotettava äärikohdan ilmaisimella 240 tapahtuvaa loppukohdan ilmaisua, ja näin ollen tunnistusprosessia viivästetään lauseen loppuun saakka. Monissa sovellutuksissa, ja nopean laitteiston yhteydessä, tämä viive ei ehkä aiheuta vaikeuksia. Muissa sovellutuksissa saattaa kuitenkin olla edullisempaa, että normalisointi aloitetaan ennen lauseen loppumista.

Sen sijaan että kertoimet normalisoitaisiin vähentämällä jokaisen kertoimen aritmeettinen keskiarvo koko lauseen ajalta, ensimmäisessä vaihtoehtoisessa suoritusmuodossa kertoimet normalisoidaan vastaavasti vähentämällä liukuva keskiarvo, joka jaksollisesti päivitetään (esimerkiksi kehys kehykseltä).

Kuvaan 9 viitaten tässä suoritusmuodossa äärikohdan ilmaisin vastaavasti ilmoittaa lauseen alkukohdan, ja sen jälkeen normalisointiprosessori 252 lukee jokaisen kertoimen osalta tämän kertoimen käsillä olevan keskiarvon $\overline{M_i}$ MFCC-kertoimen arvosta M_i ja muodostaa normalisoidun kertoimen G_i; askeltaa kehyslaskuria N; lisää kerroinarvon $\overline{M_i}$ sen hetkiseen kokonaisarvoon M_{TOT}; ja jakaa summan kehyslaskurin arvolla N, jolloin tulos talletetaan kertoimen uutena keskiarvona $\overline{M_i}$. Näin ollen jokaisen kehyksen normalisoidut kerroinarvot G_i voidaan tulostaa välittömästi.

On odotettavissa, että tämän tyyppinen liukuva keskiarvo toimii todennäköisesti heikommin kuin edullisena pidetty suoritusmuoto, koska aluksia "keskiarvoa" ei muodosteta edustavasta näytemäärästä. Toiminnassa on kuitenkin odotettavissa eräitä parannuksia verrattuna normalisoimattomiin kertoimiin. Luonnollisesti yhtä hyvin voidaan käyttää muita menetelmiä liukuvan keskiarvon laskemiseksi (esimerkiksi käyttäen aikaisempien näytteiden liikkuvaa ikkunaa, tai päivittäminen harvemmin kuin joka kehyksen kohdalla). Tämän tyyppisissä suoritusmuodoissa kerroinmuistia 251 ei tarvita.

Edellä olevissa suoritusmuodoissa on kuvattu äärikohdan osoitusta ja normalisoin-

tia saman jatkuvan lauseen yli (esimerkiksi yhden sanan yli). Jos puhujan tunnistus on tehtävä useamman erillisen sanan perusteella, niin edellä mainitussa suoritusmuodoissa selitetty prosessi voitaisiin toistaa erikseen jokaisen peräkkäisen sanan kohdalla. Puhujien erottamiseksi hyödyllistä tietoa voidaan kuitenkin löytää eri sanojen kerroinarvojen tasoista muihin verrattuna.

Vastaavasti toisessa suoritusmuodossa normalisoijalla 250 muodostettu pitkän aikavälin keskiarvo $\overline{M_i}$ muodostetaan lauseen kaikkien sanojen yli. Tämän suoritusmuodon ensimmäisessä esimerkissä tämä saavutetaan muodostamalla keskiarvo jokaisen sanan alku- ja loppukohdan kaikkien kehysten yli, aivan kuin jos sanat olisivat välittömästi peräkkäin samassa lauseessa, kun sanojen väliset puhetta sisältämättömät kehykset jätetään huomiotta.

Sama tulos saavutetaan toisessa esimerkissä, kun jokaiselle sanalle muodostetaan erilliset keskiarvot kuten edellä olevissa suoritusmuodoissa, ja kun sitten summataan keskiarvot, jotka on painotettu kulloisenkin sanan kehysten lukumäärällä, niin että muodostuu painotettu keskiarvo kaikista sanoista, ja sen jälkeen jaetaan jokaisen sanan jokainen kerroin painotetulla keskiarvolla, joka on muodostettu kaikkien sanojen avulla.

Edellä olevassa kahdessa esimerkissä jokaista eri sanaa vastaavalle keskiarvolle annettu paino vaihtelee sanan pituudesta riippuen, joka vuorostaan vaihtelee sen nopeuden mukaan, jolla puhuja lausuu sanan (muuttuja, joka ei riipu spektriominaisuuksista, jotka johtuvat tavasta jolla puhuja lausuu sanan).

Vastaavasti vaihtoehtoisessa suoritusmuodossa pitkän aikavälin keskiarvo muodostetaan muodostamalla jokaisen sanan yli keskiarvo erikseen, kuten edellä, ja muodostamalla sitten painotettu keskiarvo eri keskiarvoja käyttäen, mutta käyttäen ennalta määrättyjä painoja, jotka vastaavat esimerkiksi mallimuistiin talletetun, tätä sanaa vastaavan mallin pituutta, eikä sanan varsinaista pituutta, kuten edellisissä esimerkeissä. Tällä tavalla pienennetään riippuvuutta nopeudesta, jolla sanat lausutaan.

Vaikka eräissä tapauksissa olisi mahdollista jättää pois äärikohdan osoitin 240 ja muodostaa pitkän aikavälin keskiarvo koko puhelun kestoajan yli, niin tätä ei käytännössä yleensä pidetä edullisena, koska puhetta vailla olevien jaksojen aikana vastaanotetun signaalin taso yleensä on liian pieni luotettavan tiedon saamiseksi tietoliikennekanavan spektristä, ja lisäksi kanavan spektriä on vaikeata erottaa mahdollisesti esiintyvästä kohinasta.

Edellä selitetyissä suoritusmuodoissa on selitetty tunnistuksen käsittelylaitetta, joka

sopii kytkettäväksi tietoliikennekeskukseen. Eräässä toisessa suoritusmuodossa keksintö kuitenkin voidaan toteuttaa yksinkertaisessa laitteessa, joka on kytketty tunnusomaiseen tilaajalaitteeseen, joka vuorostaan on kytketty puhelinverkkoon. Tässä tapauksessa järjestetään analogia/digitaali-muunnosvälineet tulevan analogisen puhelinsignaalin digitoimiseksi.

Vaikka tässä on viitattu ohjelmoitavien digitaalisten signaalikäsittelylaitteiden (DSP) käyttöön, ymmärretään kuitenkin, että niiden sijasta voidaan käyttää tavanomaista yleiskäyttöistä mikroprosessoria, joka toimii riittävän nopeasti. Samaten voitaisiin käyttää asiakaskohtaista LSI-logiikkapiiriä.

Keksintöä on selitetty MFCC-kertoimiin viitaten, mutta voitaisiin myös käyttää suodatinryhmiä epätasaisella taajuusjaolla, jotka likimain noudattavat tai poikkeavat mel-taajuusjaosta. Vaikka edellä on kuvattu kolmion muotoisia ikkunoita taajuusalueella tapahtuvaa integrointia varten, on ymmärrettävä että yhtä hyvin voitaisiin käyttää muitakin ikkunan muotoja. Vaikka MFCC-arvojen laskemiseksi on kuvattu digitaalista prosessoria, niin periaatteessa sen sijasta voitaisiin käyttää useampia analogisia tai digitaalisia kaistanpäästösuodattimia, jotka vastaisivat kuvassa 5b esitettyjä kaistoja, ja ottaa tehonäytteet jokaisen suodattimen kaistalta.

Vaikka keksintö on osoitettu yllättävän edulliseksi MFCC-kertoimiin sovellettaessa, tämä ei estä soveltamasta sitä muihin edustaominaisuuksiin (edullisesti Cepstral-ominaisuuksiin).

Vaikka tässä on tarkasteltu vertailuprosessia, jossa käytetään DTW-menetelmää (Dynamic Time Warp), niin keksintöä voidaan yhtä hyvin soveltaa muunkin tyyppistä vertailuprosessia käyttävään tunnistukseen. Voidaan esimerkiksi käyttää vertailuprosessia, jossa sovelletaan HMM-menetelmää (Hidden Markov Modelling), kuten on esitetty artikkelissa S.J. Cox: Hidden Markov Models for Automatic Speech Recognition: Theory and Application; British Telecom Techonolgy Journal, Vol. 6, No. 2, huhtikuu 1988, s. 105 - 115; tai hermoverkkoja (esimerkiksi tunnettua MLP, Multilayer Perceptron; tai "itseorganisoivia" tyyppejä, joita molempia selitetään lehden British Telecom Technology Journal samassa numerossa).

Mallien muodostaminen

Yleensä esillä olevassa keksinnössä käytetään talletettua referenssimallia (malli DTW-tunnistusta varten) puhujan tai puhujien tunnistusta varten. Referenssimallien muodostamismenetelmät ovat tunnettuja, ja esillä olevan keksinnön ymmärtämiseksi riittää, että osoitetaan että jokainen malli voidaan muodostaa käsittelemällä puhujan saman sanan useampia ääntämiskertoja; digitoimalla ääntämiset: muodos-

tamalla kerrointen G_i normalisoitu joukko samalla tavalla kuin edellä selitettiin, jokaisen lausahduksen osalta; kohdistamalla lausahdukset ajallisesti käyttäen esimerkiksi DTW-prosessia; ja muodostamalla sitten lausahdusten ajallisesti kohdistettujen kerroinvektoreiden keskiarvo, niin että saadaan keskiarvoistettu kerroinvektori, joka muodostaa referenssimallin T. Tämän jälkeen tapahtuvaa tunnistusta varten käytettävä referenssimallin muodostamisprosessi määrätyllä ominaisuusjoukolla on toisin sanoen yleensä sama kuin varsinaisen ominaisuusjoukon muodostamisprosessi, jolloin tietty määrä ominaisuusjoukkoja keskiarvoistetaan referenssimallin muodostamiseksi.

Keksinnön teho

Kuvaan 10 viitaten (vasemmalla puolella oleva) kuva 10a esittää jokaisen 8 MFCC-kertoimen osalta kerroinarvon käyrän lausahduksen aikana. Jokaisessa tapauksessa on esitetty kaksi käyrää. Ne vastaavat samaa talletettua lausahdusta, joka on välitetty kahta erilaista siirtokanavaa pitkin. Havaitaan, että erityisesti toisen ja seitsemännen kertoimen osalta kanava aiheuttaa näiden kahden käyrän välillä oleellisesti vakiona pysyvän eron, joka vastaa mainitun kahden kanavan siirto-ominaisuuksien eroja vastaavilla taajuuskaistoilla.

DTW-prosessissa, kuten muissakin prosesseissa joissa osia kahdesta vertailtavasta mallista kohdistetaan ajallisesti, DTW-prosessi oleellisesti siirtää osia aaltomuodosta ajallisesti yhteensopivuuden löytämiseksi toiseen aaltomuotoon. Jos, kuten tässä, mainitut kaksi aaltomuotoa poikkeavat pystysuunnassa, niin tämä aika-akselia pitkin tapahtuva siirtäminen (eli vaakasuora siirto) johtaa ei-yhteensopivuuteen ja näin ollen tunnistuksen epäonnistumisen kasvaneeseen todennäköisyyteen tai oikean tunnistuksen pienempään todennäköisyyteen.

Oikean puoleisessa sarakkeessa olevaan kuvaan 10b viitaten on esitetty vastaavat käyrät keksinnön mukaisilla normalisoiduilla MFCC-kertoimilla. Viitaten erityisesti toiseen, kuudenteen ja seitsemänteen kertoimeen, nähdään että keskiarvon poistaminen on jokaisessa tapauksessa kohdistanut esitetyt kaksi käyrää lähemmäksi toisiaan. Kun siten puhevektoria verrataan malliin, joka on voitu johtaa erilaisen tietoliikennekanavan kautta, DTW-vertailuprosessissa tapahtuu pienemmällä todennäköisyydellä tietoliikennekanavan aiheuttama puhujan väärä tunnistus.

Kuten edellä havaittiin, reittiä (yleensä lineaarinen) puhujasta tunnistusprosessorille voi edustaa yhdistetty siirto-ominaisuus H, joka käsittää reitin peräkkäisten portaiden peräkkäisten siirtofunktioiden tulon. Tunnistusprosessorin vastaanottaman puhesignaalin jokainen spektrikomponentti käsittää taajuusalueella puhujan äänen spektrikomponenttien ja tietoliikennekanavan tai -reitin siirtofunktion spektrikom-

ponenttien tulon. Jos siten kanavan siirto-ominaisuus H olisi tunnettu, voitaisiin kanavan vaikutus puhesignaaliin poistaa jakamalla vastaanotetun signaalispektrin jokainen termi siirto-ominaisuuden H vastaavalla termillä.

Tietoliikennejärjestelmän erilaisten vaihtoehtoisten signaalireittien suuresta määrästä johtuen ei kuitenkaan suoraan voida mallintaa kanavan siirtofunktiota H. On kuitenkin havaittu, että kanavan siirtofunktio yleensä on spektrin suhteen stationaarinen (eli se ei muutu paljonkaan ajan suhteen). Jos näin ollen tarkastellaan saman spektrikomponentin aikasarjaa, siirtofunktio toimii vakiokertoimena sarjan jokaisella arvolla. Aikasarjan jokaisen komponentin geometrinen keskiarvo on tämän vuoksi tämän vakiokertoimen ja alkuperäisen aikasarjan geometrisen keskiarvon tulo. Kanavan vaikutus poistuu näin ollen, jos vastaanotetun puhesignaalin spektrin jokainen termi jaetaan sen pitkän aikavälin keskiarvolla.

Kun otetaan jokaisen spektritermin logaritmi, sen sijaan että muodostettaisiin pitkän aikavälin geometrinen keskiarvo ja suoritettaisiin jako tällä keskiarvolla, voidaan muodostaan logaritmisen spektritermin pitkän aikavälin aritmeettinen keskiarvo ja vähentää tämä aritmeettinen keskiarvo jokaisesta spektritermistä.

MFCC-kertoimia muodostettaessa on ajateltu, että ikkunointi- ja integrointivaiheissa syntyy tämän suhteen jonkinlainen muunnos, niin ettei edellä oleva analyysi kokonaisuudessaan päde MFCC-kertoimien normalisointiin, vaan ainoastaan havainnollistaa keksinnön tehoa.

Sen lisäksi että normalisointiprosessi poistaa kanavan vaikutuksen, se myös poistaa jonkin verran tietoa puheesta ja puhujasta. Tämän vuoksi voitaisiin ajatella, että se heikentää tunnistuksen tarkkuutta poistamalla dataa, jota voitaisiin käyttää kahden puhujan erottamiseksi toisistaan. Itse asiassa olemme laajojen kokeiden jälkeen yllättäen havainneet, ettei asia ole näin.

Patenttivaatimukset

1. Menetelmä puhujan tunnistamiseksi, joka käsittää vaiheet, joissa johdetaan tunnistuksen ominaisuusdata tulevasta puhesignaalista, jolloin mainittu tunnistuksen ominaisuusdata käsittää joukon kertoimia, jotka kulloinkin liittyvät puhesignaalin suuruuteen ennalta määrätyllä taajuuskaistalla; verrataan mainittua ominaisuusdataa ennalta määrättyyn puhujan referenssidataan; ja osoitetaan puhujan tunnistaminen vertailun tuloksesta riippuen; **tunnettu** siitä, että mainitut taajuuskaista jakaantuvat epätasaisesti taajuusakselia pitkin, ja että mainittujen kertoimien muodostamisvaihe sisältää vaiheen, jossa johdetaan pitkän aikavälin keskimääräinen spektrivoimakkuus; ja käsitellään ainakin yhtä mainituista kertoimista niin, että muodostuu normalisoitu kerroin, jossa mainitun pitkän aikavälin suuruus on oleellisesti pienentynyt.

2. Patenttivaatimuksen 1 mukainen menetelmä, **tunnettu** siitä, että taajuuskaistat on sijoitettu mel-taajuusasteikolle.

3. Patenttivaatimuksen 1 mukainen menetelmä, **tunnettu** siitä, että taajuuskaistat on sijoitettu lineaarisesti ennalta määrätyn rajan alapuolisilla taajuuksilla ja logaritmisesti mainitun rajan yläpuolisilla taajuuksilla.

4. Jonkin edellisen patenttivaatimuksen 1 - 3 mukainen menetelmä, **tunnettu** siitä, että vaihe, jossa muodostetaan mainitut kertoimet, sisältää vaiheen, jossa muodostetaan mainitun suuruuden logaritmi, muodostetaan pitkän aikavälin logaritminen keskiarvo, ja vähennetään logaritminen pitkän aikavälin keskiarvo logaritmisesta suuruudesta.

5. Jonkin edellisen patenttivaatimuksen 1 - 4 mukainen menetelmä, **tunnettu** siitä, että mainitussa vertailussa ominaisuusdata kohdistetaan ajallisesti referenssidataan.

6. Patenttivaatimuksen 5 mukainen menetelmä, **tunnettu** siitä, että vertailussa käytetään DTW-menetelmää (Dynamic Time Warp).

7. Jonkin edellisen patenttivaatimuksen 1 - 6 mukainen menetelmä, **tunnettu** siitä, että se lisäksi käsittää vaiheen, jossa tunnistetaan puheen alkukohta ja puheen loppukohta mainitussa puhesignaalissa; ja muodostetaan mainitun pitkän aikavälin keskiarvo mainitun alkukohdan ja mainitun loppukohdan kestoajan yli.

8. Jonkin edellisen patenttivaatimuksen 1 - 7 mukainen menetelmä, **tunnettu** siitä, että mainittu pitkän aikavälin keskiarvo on pitkän aikavälin keskiverto.

9. Jonkin edellisen patenttivaatimuksen 1 - 7 mukainen menetelmä, **tunnettu** siitä.

että mainittu pitkän aikavälin keskiarvo on liukuva keskiarvo, jota päivitetään jaksollisesti.

10. Jonkin edellisen patenttivaatimuksen mukainen menetelmä, **tunnettu** siitä, että se käsittää useamman sanan syöttämisen peräkkäin, ja mainitun pitkän aikavälin keskiarvon muodostamisen kaikkien mainittujen sanojen yli.

11. Laite puhujan tunnistamiseksi, joka käsittää välineet (210, 220, 230), jotka puhesignaalin perusteella muodostavat joukon ominaisuusdataa käsittäen kerroinjoukkojen sarjan, jolloin jokainen joukko käsittää useampia kertoimia, jotka osoittavat lyhyen aikavälin spektrisuuruuden useammalla taajuuskaistalla, ja välineet (260), joilla verrataan mainittua ominaisuusdataa ennalta määrätyn puhujan referenssidataan ja osoitetaan vastaavan puhujan tunnistus mainitun vertailun perusteella; **tunnettu** siitä, että mainitut taajuuskaistat on sijoitettu epätasaisesti taajuusakselia pitkin, ja että on järjestetty välineet (250), joilla johdetaan ainakin yhden mainitun kertoimen pitkän aikavälin keskiarvoinen spektrisuuruus, ja joilla normalisoidaan mainittu ainakin yksi kerroin tai jokainen mainittu kerroin mainitun pitkän aikavälin keskiarvon avulla.

12. Patenttivaatimuksen 11 mukainen laite, **tunnettu** siitä, että taajuuskaistat on sijoitettu mel-taajuusasteikolle.

13. Patenttivaatimuksen 11 mukainen laite, **tunnettu** siitä, että taajuuskaistat on sijoitettu lineaarisesti ennalta määrätyn rajan alapuolisilla taajuuksilla ja logaritmisesti mainitun rajan yläpuolisilla taajuuksilla.

14. Jonkin edellisen patenttivaatimuksen 11 - 13 mukainen laite, **tunnettu** siitä, että välineet (230) mainittujen kertoimien muodostamiseksi on järjestetty muodostamaan mainitun suuruuden logaritmin, muodostamaan pitkän aikavälin logaritminen keskiarvo, ja vähentämään logaritminen pitkän aikavälin keskiarvo kertoimen logaritmisesta suuruudesta.

15. Jonkin edellisen patenttivaatimuksen 11 - 14 mukainen laite, **tunnettu** siitä, että mainittu vertailuväline (260) on järjestetty ajallisesti kohdistamaan ominaisuusdatan referenssidataan.

16. Patenttivaatimuksen 15 mukainen laite, **tunnettu** siitä, että vertailuväline (260) käyttää DTW-menetelmää (Dynamic Time Warp).

17. Jonkin edellisen patenttivaatimuksen 11 - 16 mukainen laite, **tunnettu** siitä, että se lisäksi käsittää vertailuvälineet (240) alkukohdan ja loppukohdan tunnistamiseksi mainitussa puhesignaalissa, jolloin mainittu normalisointiväline (250) on jär-

jestetty johtamaan mainitun pitkän aikavälin keskiarvon lauseen kestoajan yli mainitun alkukohdan ja mainitun loppukohdan välillä.

18. Jonkin edellisen patenttivaatimuksen 11 - 17 mukainen laite, **tunnettu** siitä, että mainittu pitkän aikavälin keskiarvo on pitkän aikavälin keskiverto.

19. Jonkin edellisen patenttivaatimuksen 11 - 17 mukainen laite, **tunnettu** siitä, että mainittu pitkän aikavälin keskiarvo on liukuva keskiarvo, jota päivitetään jaksollisesti.

20. Jonkin edellisen patenttivaatimuksen 11 - 19 mukainen laite, **tunnettu** siitä, että on järjestetty keinot useamman sanan syöttämiseksi peräkkäin, ja että mainittu normalisointiväline (250) on järjestetty muodostamaan mainitun pitkän aikavälin keskiarvon kaikkien mainittujen sanojen yli.

21. Jonkin edellisen patenttivaatimuksen 11 - 20 mukainen laite, **tunnettu** siitä, että se on sovitettu kytkettäväksi puhelinverkkoon.

22. Puhelinverkko, **tunnettu** siitä, että se käsittää patenttivaatimuksen 21 mukaisen laitteen.

23. Laite äänisignaalin tunnistuskäsittelyä varten, jolloin laite käsittää välineet (210, 220, 230), jotka muodostavat tunnistusdataa, jolloin tunnistusdata käsittää joukon signaaleja, jotka kulloinkin liittyvät lyhyen aikavälin amplitudiin mainitun äänisignaalin vastaavalla taajuuskaistalla, ja välineet (260) tunnistusprosessin suorittamiseksi tunnistusdatan perusteella; **tunnettu** välineistä (250), jotka jaksollisesti muodostavat spektriamplitudin pitkän aikavälin liukuvan keskiarvon mainituilla taajuuskaistoilla tai päivittävät sitä, ja jotka käsittelevät mainittua ominaisuusdataa käyttäen mainittua pitkän aikavälin keskiarvoa vähentämään datan riippuvuutta stationaarisista spektrivaipan komponenteista.

Fig.1.

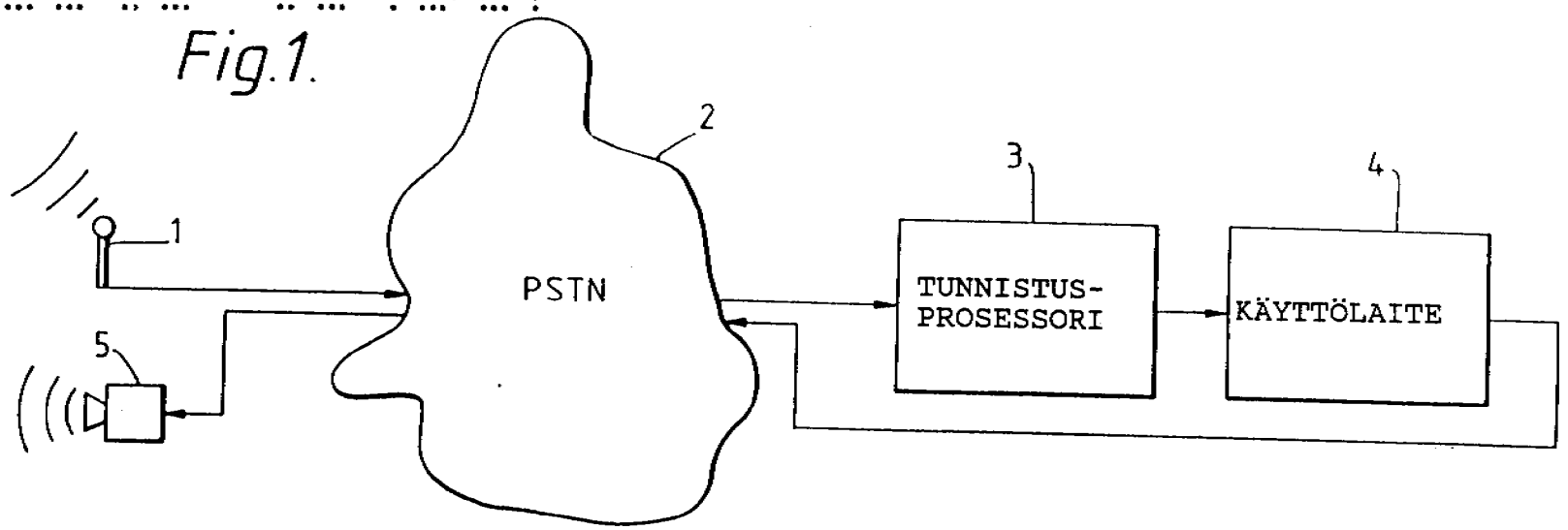

1
2
3
4
5
PSTN
TUNNISTUS-
PROSESSORI
KÄYTTÖLAITE

Fig.2.

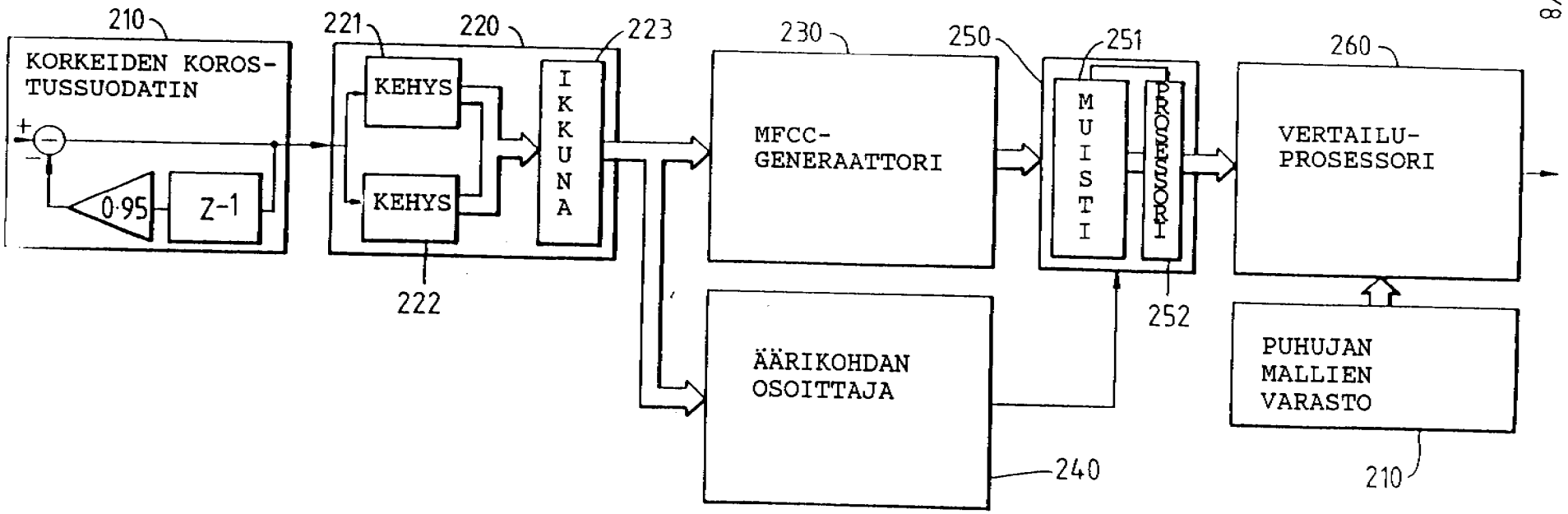

210
KORKEIDEN KOROS-
TUSSUODATIN
+
−
0·95
Z-1
221
220
223
KEHYS
KEHYS
IKKUNA
222
230
MFCC-
GENERAATTORI
250
251
MUISTI
PROSESSORI
252
260
VERTAILU-
PROSESSORI
ÄÄRIKOHDAN
OSOITTAJA
240
PUHUJAN
MALLIEN
VARASTO
210

Fig.3.

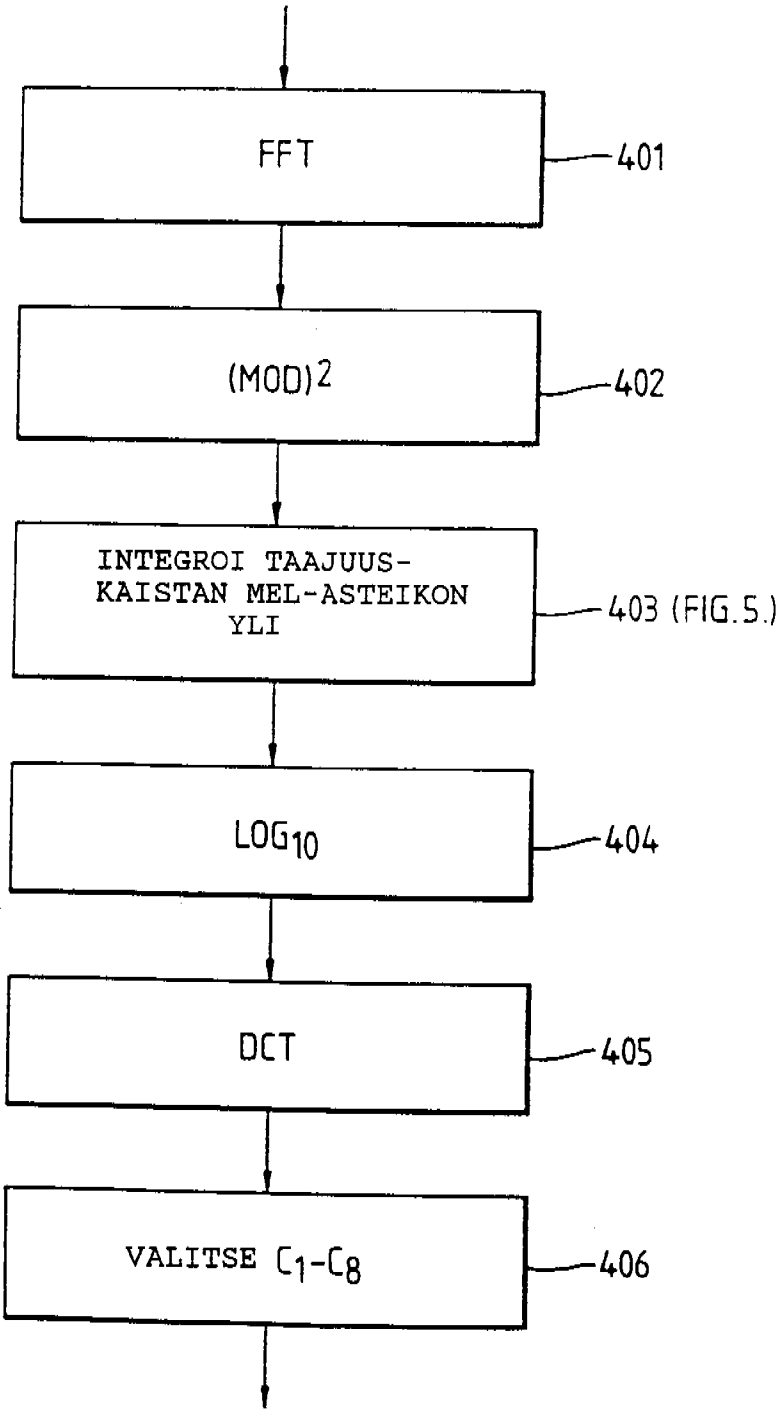

FFT
401
(MOD)2
402
INTEGROI TAAJUUS-
KAISTAN MEL-ASTEIKON
YLI
403 (FIG.5.)
LOG10
404
DCT
405
VALITSE C1-C8
406

Fig.4.

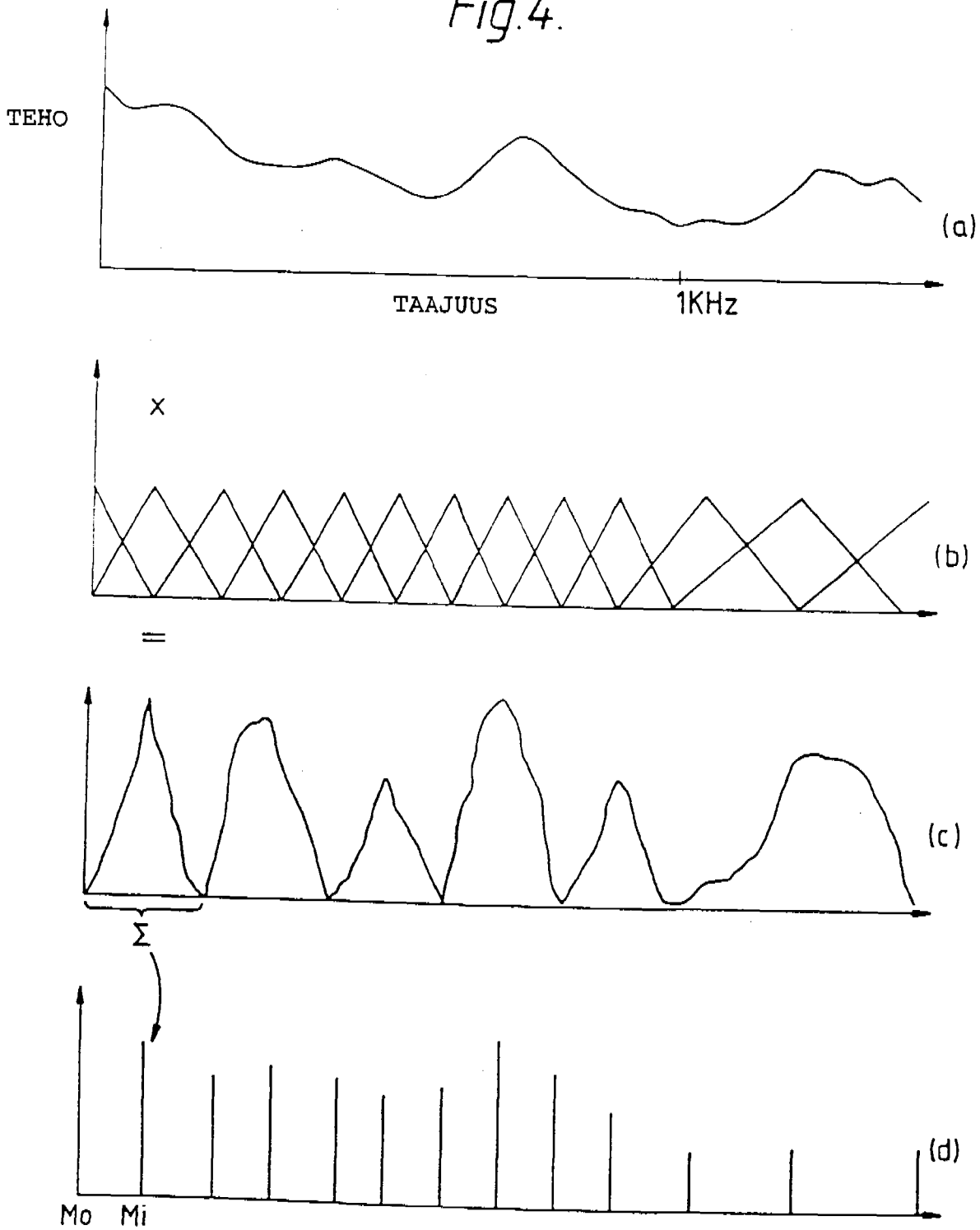


Fig. 5.

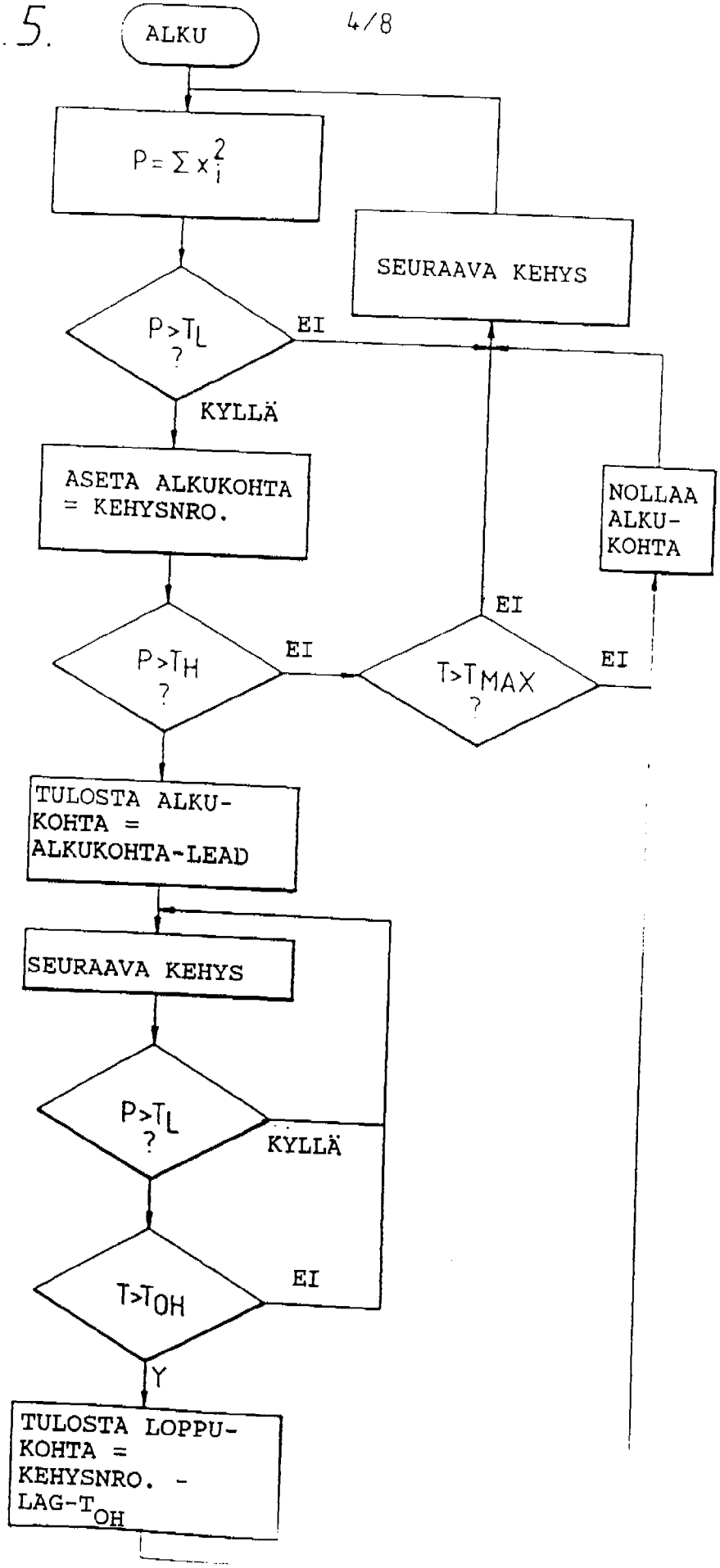

ALKU
$P = \Sigma x_i^2$
SEURAAVA KEHYS
$P > T_L$?
EI
KYLLÄ
ASETA ALKUKOHTA = KEHYSNRO.
NOLLAA ALKU-KOHTA
$P > T_H$?
EI
$T > T_{MAX}$?
EI
EI
TULOSTA ALKU-KOHTA = ALKUKOHTA-LEAD
SEURAAVA KEHYS
$P > T_L$?
KYLLÄ
$T > T_{OH}$
EI
Y
TULOSTA LOPPU-KOHTA = KEHYSNRO. - LAG-T_{OH}

Fig. 6.

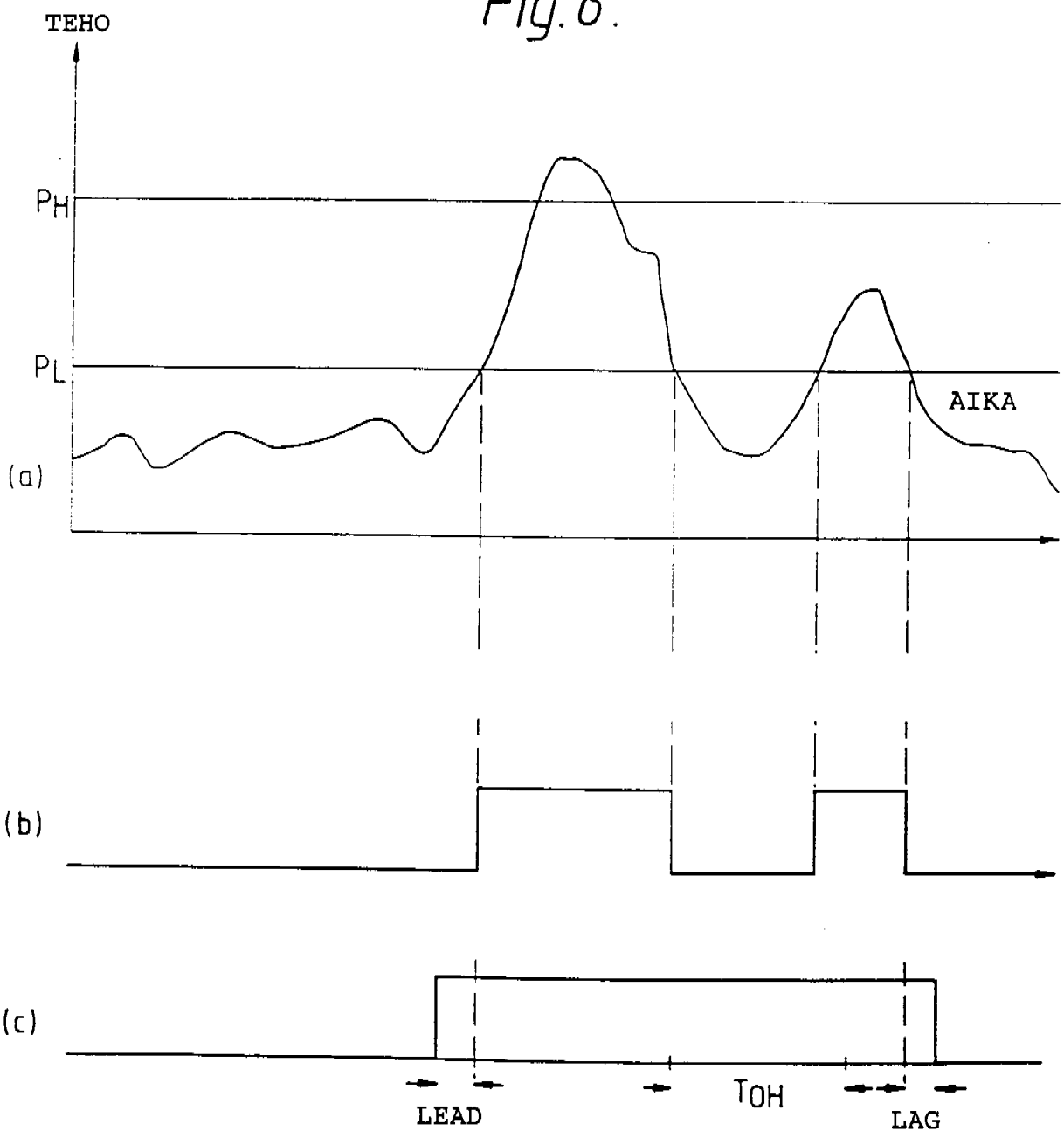


Fig. 7.

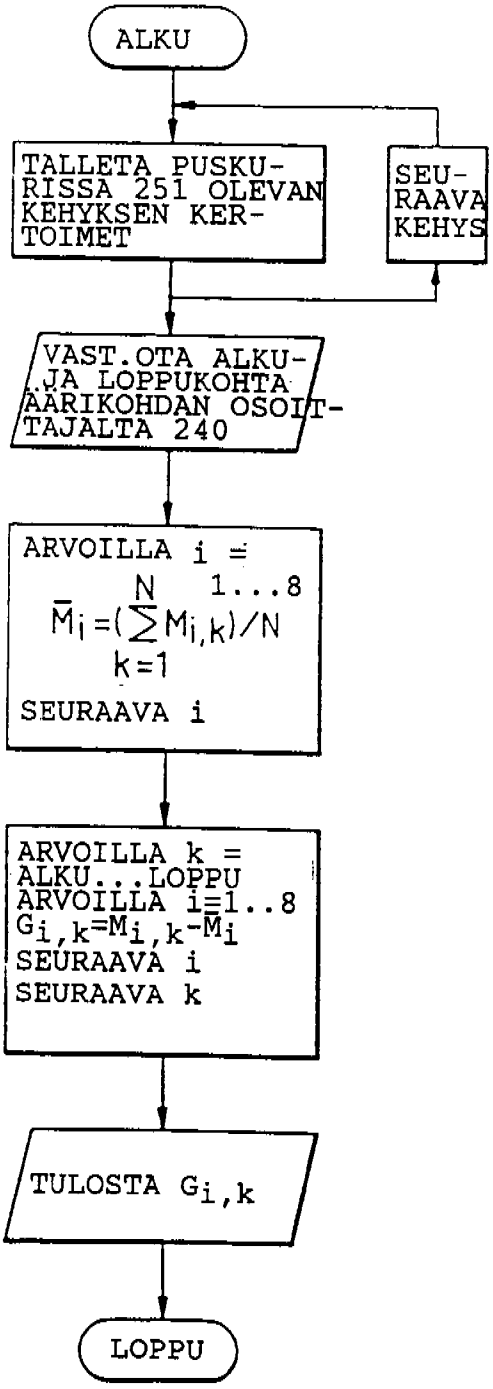

ALKU
TALLETA PUSKU-RISSA 251 OLEVAN KEHYKSEN KER-TOIMET
SEU-RAAVA KEHYS
VAST.OTA ALKU- JA LOPPUKOHTA MÄÄRIKOHDAN OSOIT-TAJALTA 240
ARVOILLA i = 1...8
$\bar{M}_i = (\sum_{k=1}^{N} M_{i,k})/N$
SEURAAVA i
ARVOILLA k = ALKU...LOPPU
ARVOILLA i=1..8
$G_{i,k} = M_{i,k} - \bar{M}_i$
SEURAAVA i
SEURAAVA k
TULOSTA $G_{i,k}$
LOPPU

Fig. 9.

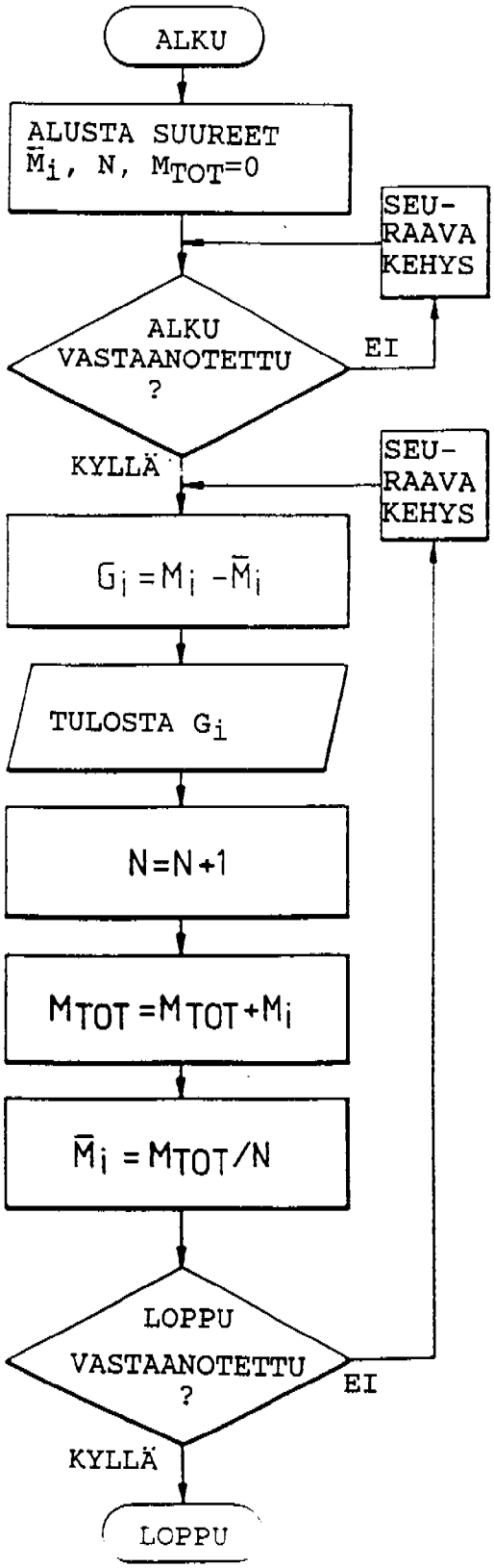

ALKU
ALUSTA SUUREET $\bar{M}_i$, N, $M_{TOT}=0$
SEU-RAAVA KEHYS
ALKU VASTAANOTETTU ?
EI
KYLLÄ
SEU-RAAVA KEHYS
$G_i = M_i - \bar{M}_i$
TULOSTA G_i
N=N+1
$M_{TOT} = M_{TOT} + M_i$
$\bar{M}_i = M_{TOT}/N$
LOPPU VASTAANOTETTU ?
EI
KYLLÄ
LOPPU

Fig. 8a.

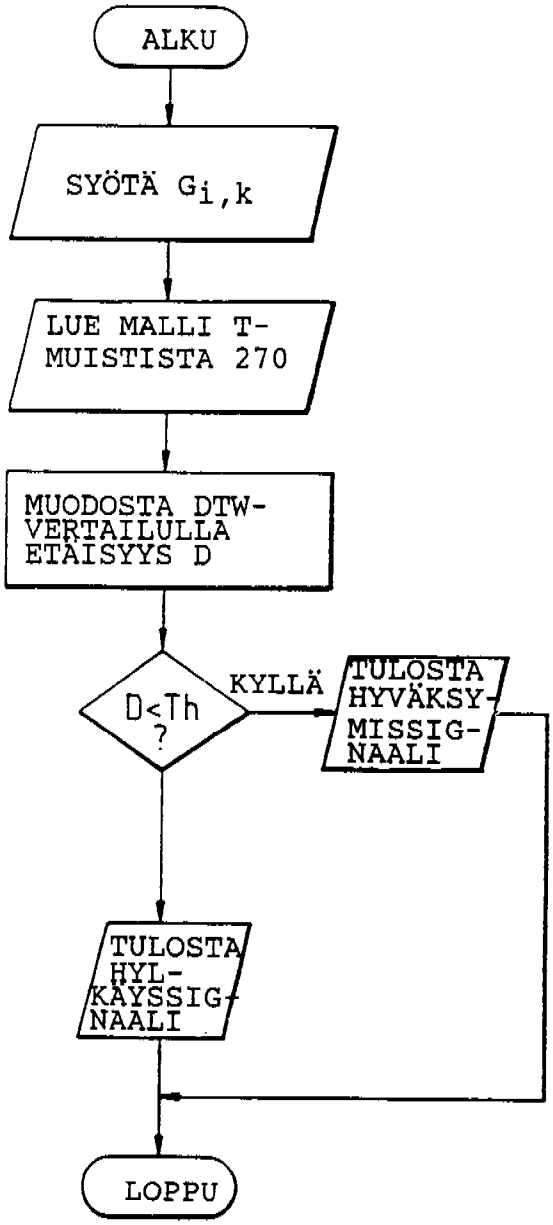

ALKU
SYÖTÄ $G_{i,k}$
LUE MALLI T-MUISTISTA 270
MUODOSTA DTW-VERTAILULLA ETÄISYYS D
D<Th ?
KYLLÄ
TULOSTA HYVÄKSYMISSIGNAALI
TULOSTA HYLKÄYSSIGNAALI
LOPPU

Fig. 8b.

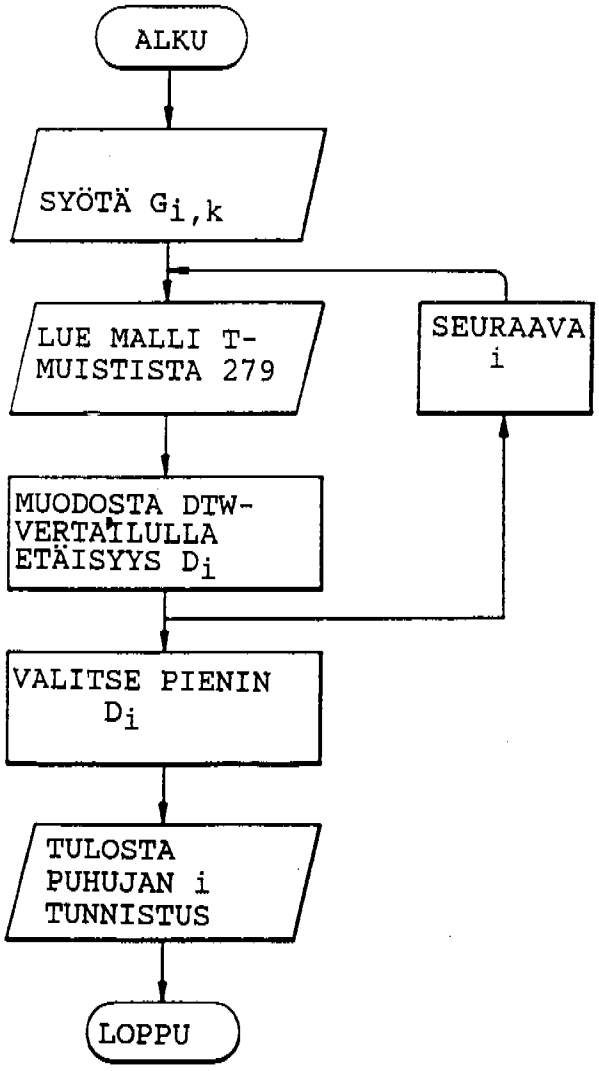
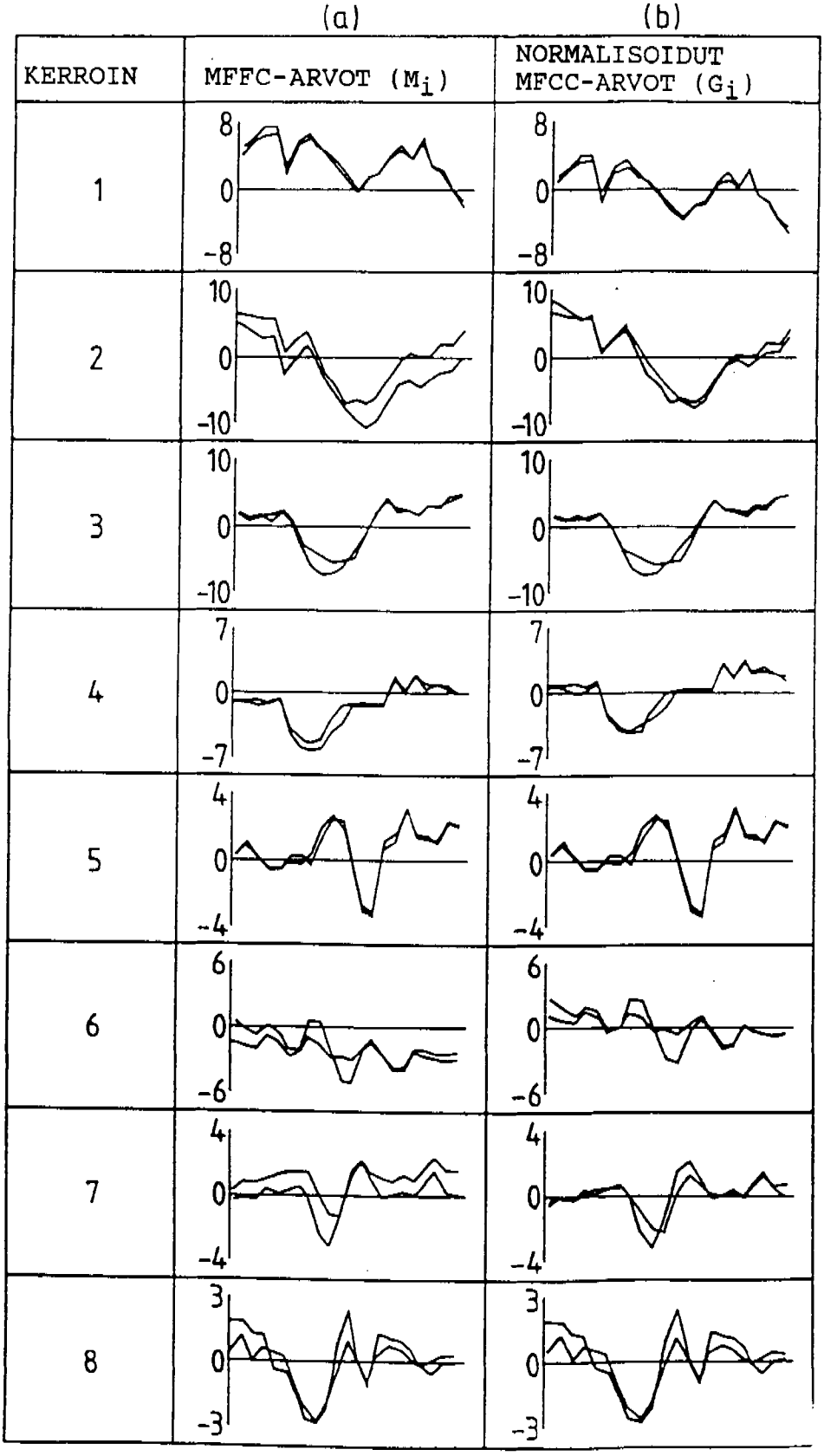

ALKU
SYÖTÄ $G_{i,k}$
LUE MALLI T-MUISTISTA 279
SEURAAVA i
MUODOSTA DTW-VERTAILULLA ETÄISYYS D_i
VALITSE PIENIN D_i
TULOSTA PUHUJAN i TUNNISTUS
LOPPU

Fig.10.


(a)
(b)
KERROIN
MFFC-ARVOT (M_i)
NORMALISOIDUT MFCC-ARVOT (G_i)
1
2
3
4
5
6
7
8
8
0
-8
10
0
-10
10
0
-10
7
0
-7
4
0
-4
6
0
-6
4
0
-4
3
0
-3