Pszichoakusztika és teremakusztika hangforrások tér-időbeni szimulált szegregációjában

2007. július 24.

*Ph.D.*disszertáció tézisei

Fodróczi Zoltán

Pázmány Péter Katolikus Egyetem
Információs Technológiai Kar

Magyar Tudományos Akadémia
Számítástechnikai és Automatizálási Kutató Intézet

Tudományos vezető:
Dr. Radványi András
az MTA doktora

Budapest, 2007

„Egyszerre csak megértettem, hogy a nyelvben, az üveggyöngyjáték nyelvében vagy legalábbis szellemében csakugyan minden mindent jelent, s minden jelkép és minden jelképváltozat nem ide vagy oda, nem egyes példákhoz, kísérletekhez, és bizonyításokhoz, hanem a középpontba, a titokba, a világ lényegébe, az őstudásba vezet."

Hermann Hesse, Az üveggyöngyjáték

Pszichoakusztika és teremakusztika hangforrások tér-időbeni szimulált szegregációjában

Összefoglaló

A konkurens hangforrások jeleinek szétválasztása régóta kutatott terület, azonban egyelőre nem állnak rendelkezésünkre olyan algoritmusok, melyek a biológiai rendszerek képességeit megközelítenék. Disszertációmban e feladat megoldásában alkalmazható két eljárással foglalkozom.

Ezek egyike a hangok fizikai jellemzők szerinti, heurisztikus algoritmusokkal történő szétválasztása. Megfigyelések igazolják, hogy az emberi hallórendszer az egyes frekvencia-komponenseket spektrális és időbeni tulajdonságaik alapján összerendeli, majd az így létrehozott csoportokat egyetlen forrásból érkező hangobjektumként kezel. Feltételezhető, hogy a hallórendszer ezen funkciója kulcs szerepet játszik a több forrásból azonos időben érkező jelek szegregációjában, ezért több számítógépes modell készült a pszichoakusztikus megfigyelésekkel azonosított funkciók mesterséges rendszerekben való alkalmazására. A közölt módszerek számítási igénye azonban ez idáig nem tette lehetővé a valós idejű alkalmazást. A dolgozatban egy celluláris hullámszámítógépen futó programkönyvtárat mutatok be, mely az emberi hallórendszer bizonyos funkcióinak hatékony megvalósítását teszi lehetővé. A közölt algoritmusokat a celluláris architektúra hardveres megvalósításából fakadó speciális követelményeknek eleget téve készítettem el, építve a már meglevő robusztus megoldásokra. A közölt programkönyvtár alkalmazásának módját egy példa alkalmazáson keresztül szemléltetem, amelyben azonos időben szimulált akusztikus térben beszélő emberek, hang alapján történő helymeghatározását valósítom meg. Bemutatom, hogy az implementált szabályok segítségével kiválasztott jelszegmenseket felhasználva a helymeghatározás hibája radikálisan csökkenthető.

A forrás-szeparációs probléma megoldásának egy másik stratégiája a források különböző térbeli elhelyezkedése alapján megvalósított szegregáció. Disszertációmban áttekintem a forrás-lokalizációs feladatok megoldását megkísérlő algoritmusokat. Rámutatok, hogy visszhangos környezetben a forrás anizotrop tulajdonságából fakadóan a hagyományos érkezési-időkülönbség becslő algoritmusok hibás eredményre vezetnek. Bemutatok egy, az akusztikus környezet hatásait figyelembe vevő forrás-lokalizáló eljárást, mely zajmentes esetben a közölt algoritmusoknál lényegesen hatékonyabban képes a forrás helyének meghatározására. Szimulációk segítségévl vizsgálom a közölt eljárás változó akusztikus körülmények között való alkalmazásának lehetőségét, illetve a számítási igényt figyelembe véve összehasonlítást végzek más korszerű forrás-lokalizáló eljárásokkal. Az alkalmazott akusztikus modell érvényességének ellenőrzését követően ajánlást adok az algoritmus gyakorlati alkalmazásának lehetőségeire.

Köszönetnyilvánítás

Mindenekelőtt szeretnék köszönetet mondani Dr. Roska Tamás professzor úrnak az MTA-SZTAKI Analogikai és Neurális Számítógépek Laboratórium vezetőjének, a Pázmány Péter Katolikus Egyetem Információs Technológia Kar dékánjának, aki türelemmel várta kutatásaim eredményét és biztosította a munkához elengedhetetlen szellemi és anyagi feltételeket.
Köszönettel tartozom Dr. Takács Györgynek, aki PhD tanulmányaim kezdetén mentorom volt. Irányításával kaptam képet az akusztika világáról. Pótolhatatlan tanácsaival egész idő alatt mellettem állt, amelyek sokat segítettek a beszéd- és a jelfeldolgozás területén a helyes irány megtalálásában.
Hálás vagyok Dr. Szolgay Péternek, aki már egyetemi hallgató koromban betekintést nyújtott a tudományos élet világába és mindvégig barátsággal támogatott.
Köszönet illeti Dr. Bércesné Dr. Novák Ágnes tanárnőt, aki a Pázmány Egyetemen tartott gyakorlati óráim előadója volt. Ágnesnek köszönhetem, hogy a tanulmányaim során teljesítendő tanítási kötelezettségnek a kutatómunkámat segítve sikerült eleget tennem. Megértése, támogatása, emberi hangja fontos volt.
Köszönettel tarozom Dr. Illényi Andrásnak a Budapesti Műszaki és Gazdaságtudományi Egyetem Távközlési és Médiainformatikai Tanszék professzorának, hogy lehetővé tette, hogy a munkámhoz nélkülözhetetlen méréseket és kísérleteket elvégezzem a tanszék kezelésében levő Békésy György Akusztikai Laboratórium egyedülálló süketszobájában.
A Ph.D hallgató nyugodt, ám a felszín alatt viszontagságokkal és kételyekkel teli hétköznapjainak elviselésében megkérdőjelezhetetlen érdeme van az analogikai laborban, illetve a Pázmány Egyetemen dolgozó kollégáimnak: Kis Attilának, Wágner Róbertnek, Jónás Péternek, Benedek Csabának, Hegyi Barnabásnak, Havas Lászlónak, Vásárhelyi Gábornak, Lázár Annának, Bárdi Tamásnak, Harczos Tamásnak és Feldhoffer Gergőnek.
A németországi Fraunhofer intézetben eltöltött szemeszter során szellemi és erkölcsi támogatást kaptam Dr. Frank Klefenztől, Kátai Andrástól, Stephan Wernertől és Wolfgang Köstritzertől.
Publikációim angolságának tökéletesítése miatt elismeréssel adózom Nagy Éva Nórának, János Kornnak és Péri Mártonnak.
Tanulmányaim során a Magyar Tudományos Akadémia Számítástechnikai és Automatizálási Kutató Intézetének támogatása nélkülözhetetlen volt.
A témavezetőm, a családom, a barátaim érdemeinek kifejezésére jelen keretek közt nem vállalkozom.

Tartalomjegyzék

A dolgozatban használt jelölések

$\mathcal{L}(\mathrm{l})$	a hangforrás l helyen való elhelyezkedésének valószínűsége
$\mathrm{p}^{\mathrm{RM}}_{s,\varphi,\theta}$	az (s,φ,θ) akusztikus konfiguráció becsült visszhanghatás-térképe
$c_{u,u}(0)'_+$	az auto-korrelációs függvény 0 helyen vett jobb oldali deriváltja
$c_{u,u}(0)'_-$	az auto-korrelációs függvény 0 helyen vett bal oldali deriváltja
$c_{x_i,x_j\setminus(f,g)}(\tau_f-\tau_g)'_+$	a kereszt-korrelációs függvény f és g visszaverődési utak hatása nélküli alakjának $(\tau_f-\tau_g)$ helyen számolt jobb oldali deriváltja
$c_{x_i,x_j\setminus(f,g)}(\tau_f-\tau_g)'_-$	a kereszt-korrelációs függvény f és g visszaverődési utak hatása nélküli alakjának $(\tau_f-\tau_g)$ helyen számolt bal oldali deriváltja
$p_{s,x_i,x_j}(k)$	a forrás s pontban való elhelyezése esetén a c_{x_i,x_j} lokális maximumait jósló függvény
$T_{TDOA}(.)$	$C \to \mathbb{S}_{TDOA}$ transzformáció
$\mathcal{L}_{\max}$	az összegzett korrelációs térkép maximuma
$\alpha(.,.)$	a hang csillapítása valamely terjedési úton
$\beta(r)$	az r visszaverő felület frekvenciától és beesési szögtől független abszorpciós koefficiense
λ	a periódikus rezgés hullámhossza
$\psi_{i,j}$	az általános kereszt-korrelációs függvény kiszámításához használt súlyfüggvény
$\tau_{q,i}$	az $i.$ mikrofon q pontra fókuszálását végző nyalábirányító késleltetés
τ_p	a hang számára a p terjedési út megtételéhez szükséges idő
θ	vertikális irányszög

φ	horizontális irányszög
$\widehat{\mathcal{L}}$	az összegzett korrelációs térkép lokális maximum helyeinek halmaza
$\widehat{\widehat{\mathcal{L}}}$	az összegzett korrelációs térkép adott limit feletti (T_r) lokális maximum helyeinek halmaza
$\widehat{\widehat{p^{RM}_{s,\varphi,\theta}}}$	a becsült visszhanghatás-térkép adott limitet (T_r) meghaladó lokális maximumainak halmaza
$\widehat{p^{RM}_{s,\varphi,\theta}}$	a becsült visszhanghatás-térkép lokális maximumainak halmaza
$\widehat{s}$	a forrás feltételezett helye
$\xi_m(\varphi,\theta)$	a mikrofonok iránykarakterisztikája
$\xi_s(\varphi,\theta)$	az s hangforrás iránykarakterisztikája
$A(ij;kl)$	a $C(i,j)$ cella visszacsatoló template-jének (k,l) eleme
$B(ij;kl)$	a $C(i,j)$ cella előrecsatoló template-jének (k,l) eleme
C	a hangforrás lehetséges térbeli pozícióinak halmaza
C_A	a „lehetséges konfigurációk” rendezett-hármasait tartalmazó halmaz $(s,\varphi,\theta) \in C_A$
$c_{u,u}$	az u jel auto-korrelációs függvénye
$c_{x_i,x_j\setminus(f,g)}(k)$	a kereszt-korrelációs függvény az f és g visszaverődési utak hatása nélkül
$c_{x_i,x_j}(k)$	x_i és az x_j alapján számított kereszt-korrelációs függvény értéke a k helyen
D	a mikrofonok fizikai távolságából adódó legnagyobb lehetséges érkezési időkülönbség
d_p	a p terjedési út hossza
f,g,p,q	tetszőleges hang terjedési utak $f,p \in P_i$, $g,q \in P_j$
f_C	a becsült visszhanghatás-térképek és a megfigyelés alapján kiválasztott „lehetséges konfigurációk” halmaza
$G_{\cdot}(\omega)$	az általános kereszt-korrelációs függvény számításánál alkalmazott szűrő
m_i	az i. mikrofon pozíciója
N	a forrás-lokalizációra használt mikrofonok száma, illetve a 3. fejezetben a CNN tömb szélessége
$P(t)$	a tér egy pontjában mérhető eredő nyomás
$p(t)$	hangnyomás

P_0	az atmoszferikus nyomás
P_i	terjedési utak halmaza, melyek a forrástól az i. mikrofonig terjednek
$P_{cg}(M)$	tetszőleges M ponthalmaz súlypontja
$P_{icg}(M)$	az M ponthalmaz inverz súlypontja
$p_{s,\varphi,\theta,x_i,x_j}(k)$	az s pontban elhelyezett forrás φ és θ horizontális és vertikális iránya esetén az i és j mikrofonok által rögzített jelekből számolt kereszt-korrelációs függvény lokális maximum becslő függvénye
$p_{x_i,x_j}(k)$	a c_{x_i,x_j} lokális maximum helyeit jósló függvény. A k helyén levő lokális maximum mérete, illetve lokális maximum kialakulásának valószínűsége.
R_p	$p \in P_i$ tetszőleges terjedési út során érintett visszaverő felületek listája
R_{x_i,x_j}	a kereszt-korrelációs függvény definiciója
s	a hangforrás térbeli pozíciója
$S_r(i,j)$	az r-távolságban levő, az (i,j) elemmel összeköttetésben levő cellák halmaza.
T_r	a legkisebb figyelembe vett visszhanghatás érték
$u(k,q)$	a q pontra fókuszlált mikrofontömb által rögzített jel a k. időpillanatban
$u(t)$	a forrás által kibocsátott jel időfüggvénye
$w_p(k)$	a p-ik ablak k-ik időpillanatban felvett értéke ($k = 1...W$)
$X_i(\omega)$	x_i frekvencia-tartománybeli megfelelője
$x_i(t)$	az i. mikrofon által rögzített jel
$\mathbb{S}_{TDOA}$	az érkezési-időkülönbségek tere
$\mathrm{p}^{\mathrm{RM}}_{s,\varphi,\theta}(\mathrm{l})$	az $l \in C$ pontban a becsült visszhanghatás értéke
c	a hang terjedésének sebessége levegőben szobahőmérsékleten (c=344m/s).
C(i,j)	a CNN hálózat i-edik sorának j-edik oszlopában lévő cellát jelöli
f	a periódikus rezgés frekvenciája
M(m)	az M ponthalmazhoz m helyen hozzárendelt érték, ami praktikusan $\widehat{\widehat{p^{RM}_{z,\varphi,\theta}}}(m)$-el, illetve $\widehat{\widehat{\pounds(\mathrm{m})}}$-el egyenlő
O	a spektrogram készítésnél alkalmazott ablakok átfedése
r	3. fejezetben a CNN hálózat összeköttetési távolsága, tetszöleges visszaverő felület azonosítója egyébként (5. és 6. fejezetek)

s(i)	a digitalizált hang i. időpillanatban felvett értéke
t	folytonos időváltozó
u	a CNN cella bemenete
W	a spektrogram készítésnél, illetve a korreláció számításnál használt ablakok mérete
x	a CNN cella állapotváltozója
y	a CNN cella kimenete

1. FEJEZET

Bevezetés

A 21. század hajnalán jellemzően ugyannak a százegynéhány billentyűnek az egymás utáni leütésével kommunikálunk számítógépeinkkel, mint 1948-ban a Binac számítógép felhasználói[1], ugyanannak a kétdimenziós helymeghatározó eszköznek a segítségével jelöljük ki a figyelmünk tárgyát képező információkat, mint 1964-ben Douglas Engelbar[2], telefonjainkat továbbra is miniatűr billentyűzetek segítségével irányítjuk, holott kézenfekvő elvárásunk, a tudományos fantasztikus művek egyik legalapvetőbb víziója, hogy eszközeinket hangutasításokkal vezéreljük, hogy a mesterséges rendszerek képesek legyenek a hanghullámok szállította információt értelmezni, feldolgozni.

A több évtizedes kutatómunka eredményeként ma már léteznek olyan algoritmusok, melyek néhány százalékos hibával, az emberi hallgatók teljesítményét megközelítően képesek zaj és visszhangmentes körülmények között rögzített felvételek alapján beszéd-, hangesemény-, illetve hangfelismerést végezni. A mesterséges rendszerek hatékonysága azonban az adott feladathoz szigorúan nem kapcsolodó jelekkel (zajokkal) terhelt bemeneti információk feldolgozása esetén - ilyenek lehetnek például az utcáról beszűrődő hangok, zene, illetve egyéb, a mindennapi életben állandóan előfoduló háttérzajok - exponenciálisan romlik.

Mivel az emberi hallgatók teljesítménye hasonló körülmények között gyakorlatilag változatlan marad [1], felvetődik a kérdés, hogy melyek azok a feldolgozási lépések, amelyek döntően befolyásolják a mesterséges és a természetes rendszerek teljesítménye közötti különbséget.

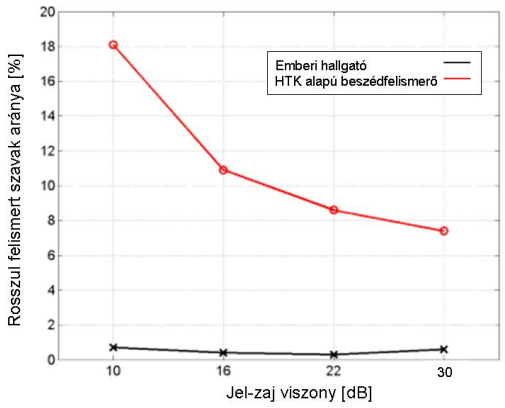
Az egyik ilyen feldolgozási lépés a fülekbe jutó hangok szemantikus elemzése, vagyis az, hogy a konkurens források jeleinek keverékét tartalmazó információt bizonyos mechanizmusok segítségével szétválasztjuk. Az elemzés eredményeként az egy időben sugárzó források jelei, elkülönítve állnak a további, figyelmünk által befolyásolt feldolgozási lépések rendelkezésére. A fenti probléma koktél parti effektusként[3] ismert az irodalomban.

A megoldandó probléma formálisan tehát *"több hangforrás jelének keverékéből álló jelfolyam, a hangforrások által kibocsátott összetevőkre való felbontása"*-ként fogalmazható meg. A fenti probléma megoldását célzó munkákat hagyományosan a forrás-szeparáció témakörébe sorolják. Dolgozatom témája az ide tartozó munkák áttekintése és az általam kidolgozott új eljárások bemutatása.

[1] A Binac volt az első számítógép, melyre a lyukkártya olvasón kívül manuálisan is lehetett adatot rögzíteni egy a géphez csatlakoztatott írógép billentyűzete segítségével [http://inventors.about.com]

[2] Az egér kifejlesztője[http://inventors.about.com]

[3] A jelenségre 'coctail party effect'-ként először Colin Cherry hivatkozott 1953-ban [2], jóllehet a probléma gyökerei alapvetően a repülés irányítók által fogadott rádió üzenetekkel kapcsolatos nehézségekre vezetnek vissza, amelyekben több pilóta hangjának keveréke volt hallható.



1.1. ábra. Egy korszerű beszédfelismerő rendszer teljesítményének alakulása különböző jel-zaj viszonyok között, összehasonlítva az emberi hallgatók teljesítményével [1].

A forrás-szeparációval foglalkozó dolgozatok a probléma megközelítésének filozófiáját tekintve három fő csoportba sorolhatóak. Ezek egyike, a több szenzor által rögzített jelfolyam statisztikai jellemzők szerinti szétválasztását célozza[4]. Dolgozatomban, az ebbe a csoportba tartozó munkákkal részletesebben nem foglalkozom, mivel a feladat megoldása a rögzített jelek különböző statisztikai paramétereire vonatkozó feltételek megléte esetén lehetséges, ezek pedig a gyakorlatban jobbára nem biztosíthatóak.

A kijelölt feladat megoldásának egyik elterjedt módja a hangok fizikai jellemzők szerinti, heurisztikus algoritmusokkal történő szétválasztása. Ennek megfelelően a 2. fejezet a hangok spektrális és időbeni tulajdonságok alapján történő szétválasztását végző mechanizmusok tárgyalását tartalmazza. A fejezetben bemutatom a pszichoakusztikai megfigyelésekből ismert csoportosítási szabályokat, valamint kitérek az adott szabály kialakulását magyarázó fizikai törvényszerűségekre. A már létező, *hallási jelenet elemzést*[5] megvalósító modellek bemutatását követően a 3. fejezetben ismertetem a Celluláris Hullámszámítás elméletének és a felhasználás gyakorlatának alapjait, majd a 4. fejezetben rátérek a 1. tézispont, az Celluláris Hullámszámítógépen megvalósított *hallási jelenet elemzés* könyvtár bemutatására. A fejezetet alkalmazási példa zárja.

Forrás-szeparációt megvalósító munkák egy csoportjának tekintem az akusztikus források helyének meghatározást végző algoritmusokat, jóllehet ezek egy része közvetlenül nem valósítja meg a jelek keverékének szétválasztását. Mivel azonban az ide tartozó módszerek egyike a hang sugárzók különböző térbeli elhelyezkedése révén képes a források jeleinek szegregációjára - és egyidejüleg a forrás helyének meghatározására -, a felvetett probléma megoldására alkalmasnak tekinthető algoritmusok. Érdemes megjegyezni, hogy az itt kitűzött feladat megoldásán túl, a forrás-lokalizáló algoritmusok elterjedt felhaszálási területe a bizontsági, illetve video konferencia rendszerek, ahol beszélők, illetve más zajforrások nyomon követése révén, nehezen helyettesíthető attenció-szelekciós

[4] Blind Source Separation.

[5] Az angol nyelvű irodalomban Auditory Scene Analysis-ként (ASA) ismert a fenti szabályszerűségek megfigyelésével foglalkozó tudományág. Winkler István [3] munkája nyomán a *hallási jelenet elemzés* magyar megfelelőt használom.

mechanizmusként szolgálnak. A fentieknek nyomán, az 5. fejezetben áttekintem a különböző hangforrás lokalizáló algoritmusokat, valamint külön fejezetben ismertetem az általam kidolgozott 2. tézispontként megfogalmazott, a környezet akusztikai adottságait integráló forrás-lokalizáló eljárást, mely az irodalomban közölt munkák közül egyedüliként alkalmas anizotrop források visszhangos környezetben való helyének meghatározására. A fejezet további részében kitérek az alkalmazott akusztikus modell érvényességének vizsgálatára, illetve összehasonlítom a bemutatott módszer hatékonyságát, más korszerű forrás-lokalizáló eljárássokkal.

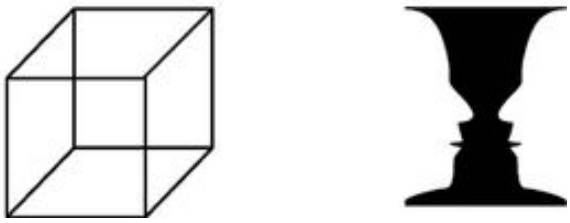
A dolgozat utolsó fejezetében az eredmények rövid összefoglalása, valamint a doktori tanulmányaim során szerzett tapasztalatok alapján, a téma lehetséges további kutatási irányaira vonatkozó gondolataim olvashatóak.

2. FEJEZET

Az emberi hallórendszer vizsgálatának módjai és mesterséges modelljei

2.1. A hallórendszer funkcióinak pszichoakusztikai módszerekkel történő azonosítása

1912-ben Max Wertheimer két kollégájával, Wolfgang Köhlerrel valamint Kurt Koffkaval egy pszichológusok számára készített jegyzetben [4] arra a kérdésre keresték a választ, hogy az állóképek meghatározott gyorsaságú vetítése miért kelti a megfigyelőben a mozgás benyomását. A válasz keresése közben egy új pszichológiai irányzat született, a Gestalt iskola[1], mely az észlelés folyamatának átfogó elméletévé nőtte ki magát. Az agyat hollisztikus egészként kezeli, mely a külső ingerek alapján, tanult vagy veleszületett mintákat felhasználva alkotja meg az érzékelt ingerekhez leginkább illeszkedő külvilág- modellt.



2.1. ábra. Multistabilitás. Az érzékelés befolyásolható. A bal oldalon hajlamosak vagyunk egy háromdimenziós kocka oldalait látni, a szimpla vonal háló helyett. A jobboldali kép egy váza, vagy két egymásfelé néző arc? ([5] forrás http://en.wikipedia.org)

A Gestalt iskola eredményeinek mesterséges rendszerekben való alkalmazási módja az egymástól

[1] die Gestalt: alak, forma, alakulás

alapvetően független primitív mintáknak (egyenesek, vonalak) az érzékelés során létrejövő csoportokba, objektumokba (kocka) való rendezése.

A Gestalt iskola megfigyeléseit a vizuális érzékeléssel kapcsolatos kutatásokban már az 1960-as évektől kezdődően vizsgálták, ezzel szemben a hang érzékelésre vonatkozó következtetések összefoglalása meglehetősen későn, 1992-ben történt meg, Albert S. Bregman [6] összefoglaló művének megjelenésével. Bregman munkájában összefoglalja, valamint egységes keretbe rendezi az addig megismert pszichokausztikai kísérleteket, melyek a hallórendszerünk által egymástól elkülöníthető érzékelési egységek (érzékelési primitívek) meglétét igazolják. Kifejti, hogy vélhetően ezen érzékelési egységek alapján történik meg a hangtér összetartozó eseményekre bontása, hierarchikusan rendezett funkciók iteratív alkalmazásának eredményeként. A hierarchia legalsó szegmensében helyezkednek el a pusztán szenzoros információk kiértékelése révén ható, úgynevezett adatvezérelt csoportosítási szabályok[2], míg a hierarchia csúcsát a tanult, illetve kontextus-függő, sémavezérelt[3] csoportosítási mechanizmusok alkotják, melyek eredménye a hangtér kognitív leképezése, azaz későbbi feldolgozási lépések igényeihez igazodó felbontással reprezentált, egy adott forrás jelét tartalmazó érzékelési folyam[4] vagy hangobjektum[5].

2.1.1. Adatvezérelt csoportosítási szabályok

A hallórendszer vizsgálata során végzett pszichokusztikai kísérletekkel sikerült néhány primitív, csak a bemeneti adatoktól függő csoportosítási szabályt azonosítani [6]. A mesterségesen előállított szinuszos hangokon végzett kísérletek eredményeinek csoportosítási szabályokba való rendezése azonban nem egyértelmű. A szabályok által definiált csoportok gyakran nem diszjunktak, a témával foglalkozó dolgozatokban nem egységes az egyes szabályok különféle bemeneti jelek esetén való értelmezésének módja, illetve tapasztalható némi bizonytalanság az elnevezéseket illetően is. Dolgozatomban igyekeszem a *hallási jelenet elemzés* körében tárgyalt mechanizmusokat az irodalomban található legspecifikusabb szempontrendszer szerint áttekinteni, melynek eredményei az alábbiak.

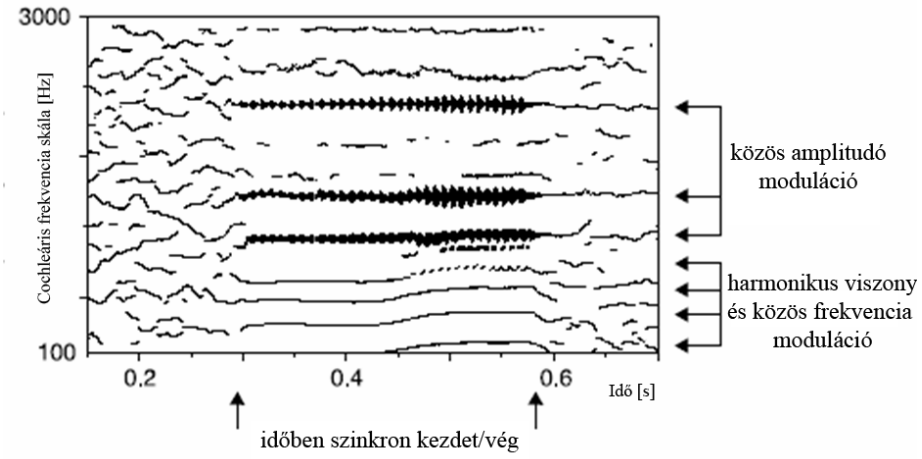
Azonos időben kezdődő/végződő komponensek csoportosítása[6]

A természetes fizikai folyamatok által keltett hangok esetén, a kisugárzott energia minden frekvenciatartományban azonos időben jelenik meg. Vélhetően evolúciós előnyt jelentett tehát a fenti törvényszerűségből adódó lehetőségek kihasználása, azaz az időben szinkron kezdődő különböző frekvenciájú komponensek egy hangobjektumként való azonosítása (lásd 2.2. ábra). Az azonos időben megszűnő különböző frekvenciájú komponensek hasonló elv alapján csoportosíthatóak egy hangobjektumba tartózókként, jóllehet ezen csoportosítási szabály szerepe jóval kevésbe hangsúlyos.

Közös sors - Azonos frekvencia-, illetve amplitúdó-moduláció [7]

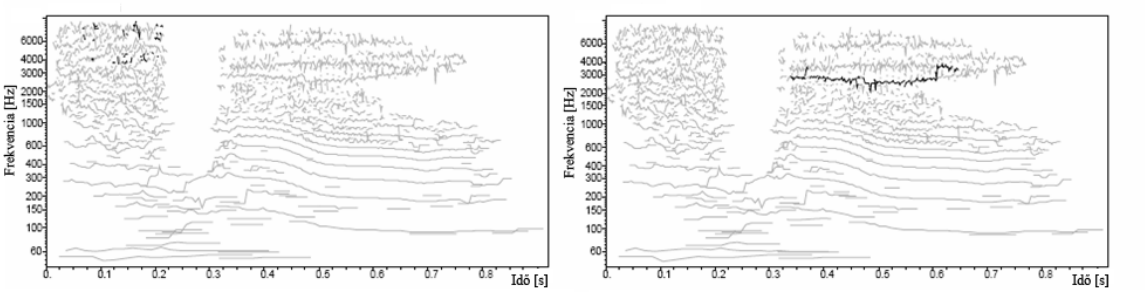
Az azonos frekvenciával, illetve amplitúdóval modulált, a 2.2. ábrán látható viszonyban levő, egymással nem átfedő frekvencia-komponensek egyes elemeinek megkülönböztetése az emberi hallgatók számára meglepően nehéz feladat, mivel egyetlen összetett hang jelenlétét tapasztaljuk. A jelenség legvalószínűbb magyarázata, hogy hallórendszerünk alkalmazkodott a természetben előforduló

[2] data driven rules
[3] schema driven
[4] stream
[5] Egyetlen forrásból eredőként érzékelt komponensek összessége.
[6] Common Onset/Offset rule
[7] Common Fate



2.2. ábra. Példa a közös kezdet/vég, a közös sors, valamint a harmónikus viszony csoportosítási szabályokra. (forrás [7, 8])

frekvencia-, illetve amplitúdó-modulált hangokat kibocsátó folyamatok észlelésére. Ezek közös jellemzője, hogy a kisugárzott energia minden komponensében időben szinkron modulált, azaz például beszédhang esetén, a hang erejének, illetve magasságának változása a hangképző szervek által keltett összes komponens azonos arányban és értelemben való változását eredményezi.



2.3. ábra. A folytonosság (bal) és a közelség (jobb) csoportosítási szabályok értelmezését segítő ábra. Szürké vonalak jelzik a Q-színusz transzformáció eredményét, míg a kiemelt részek a csoportosítási szabályok által azonosított komponensek. (forrás [9, 10])

Folytonosság[8]

Az egyazon forrásból származó, csak meghatározott frekvenciákat tartalmazó hangok rövid időre megszünhetnek, majd a megszünés pillanatában aktuális frekvenciához közel ismét újra kezdődhetnek. Az emberi hallórendszer a frekvenciában és időben egymáshoz közeli komponenseket azonos

[8] Continuity

hangobjektumokként észleli (lásd 2.3. ábra).

Közelség[9]

Az egymáshoz frekvenciában és időben közeli, a 2.3. ábrán látható, azonos energiájú, rövid idejű energia tüskék csoportosítására szolgáló szabály. A létézésére vonatkozó pszichoakusztikai megfigyelések javarészt megegyeznek a folytonosság szabálynál említettekkel, azonban ezúttal a zajszerű energia tüskék csoportosításaként értelmezett ezért külön tárgyalt. Jelentősége a beszédben előforduló zöngés mássalhangzók okozta komponensek csoportosításában elsődleges.

Harmonikusság[10]

Ez az egyik legfontosabb és csak a hang észlelésben kialakult csoportosítási szabály, melynek nem található megfelelője a vizuális érzékelésben. Azokat a komponenseket, melyek frekvenciái egymásnak egész számú többszörösei - harmonikusai - egyetlen komplex hangként érzékeljük. Az egyes komponensek egymáshoz viszonyított aránya és erőssége adja az adott hang hangszínként érzékelt tulajdonságát. Megfigyelések igazolják, hogy a harmonikus kapcsolatban levő komponensek észlelésben való csoportosításának valószínűsége fordítottan arányos az egyes komponensek frekvencia távolságával (azaz az első és a harmadik harmonikus együttes észlelése valószínűbb mint az első és az ötödik harmonikusé). További tény, hogy a köztes harmonikusok hiánya ugyancsak csökkenti a komponensek csoportos észlelésének valószínűségét. A megfigyelt csoportosítási szabály kialakulása ugyancsak a természetben előforduló hangok sajátosságaira vezethető vissza. A fizikai rezgő rendszerek (pl. húrok vagy a hangszalagok) általában nem egyetlen frekvencián rezegnek, hanem a modális oszcilláció révén harmonikus frekvenciákon is.

2.1.2. Sémavezérelt csoportosítás

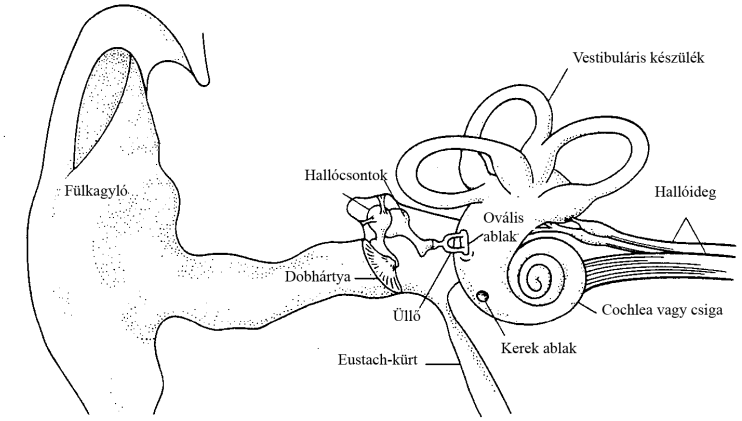
Az eddig bemutatott adatvezérelt mechanizmusokon túl, melyek csak a külvilágból érkező szenzoros információk alapján valósítják meg az egyes hangkomponensek csoportokba sorolását, azt feltételezzük, hogy a hallórendszer tanult mintákat is felhasznál a hang komponensek összetartozó hangobjektumokká való szervezésében. Ennek egyik bizonyítéka egy 1970-ben végzett kísérlet [11], amelyben értelmes mondatok egyes szavainak bizonyos fonémái helyett zajt sugároztak a kísérleti alanyok fülébe. Az alanyok, a mondat értelmétől függően "hallották" a zaj helyére a megfelelő fonémákat, tehát tanult mintáik, illetve felsőbb kognitív folyamatok révén, a hallórendszer a mondatok értelme alapján transzformálta a bejövő hibás információt konzisztens hallási élménnyé.

Napjainkban, az imént említett felsőbb mechanizmusok működéséről egyelőre keveset tudunk, az azonban bizonyosnak látszik, hogy a hallórendszer, a látási illúzióknál bemutatott példákhoz hasonlóan képes az előző fejezetben bemutatott primitív csoportosítási szempontok eredményeit figyelembe véve megkeresni a külvilágból érkező információk eredetéért felelős legvalószínűbb modellt. Az agyunk által, az észlelés során előállított komplex hallásélmény tehát egy adaptív-iteratív folyamat hatására alakul ki, melyben a primitív csoportosítási mechanizmusok által szolgáltatott redundáns, átfedő, nem ortogonális csoportok egymással versengve adják a hangtérből érkező információk dekompozicióját.

[9] Proximity
[10] Harmonicity

2.2. A hallórendszer funkcionális anatómiája

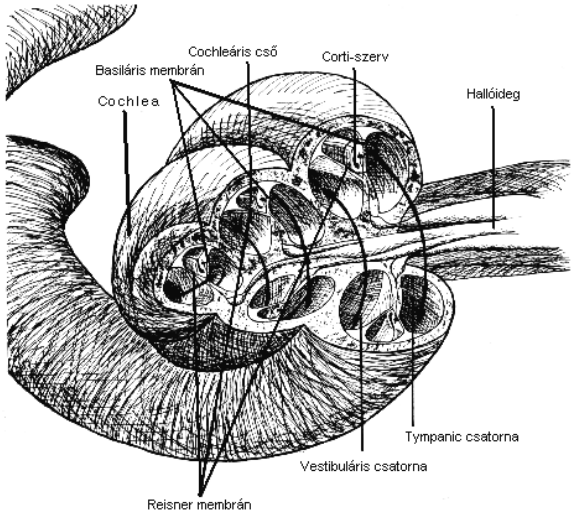
A biológiai rendszerek működésének, pszichológiai módszerekkel történő vizsgálatakor, az elsődleges cél a komplex rendszer egyes funkcióinak, illetve ezen funkciók sajátosságainak azonosítása. Mérnöki terminológiával, a komplex rendszer viselkedéséért felelős „szoftver" működési módjainak megismerése az elsődleges feladat. A kijelölt probléma megoldása érdekében szükséges azonban a rendszert megvalósító „hardver" elemekről megszerezhető információkat is figyelembe venni. Az áttekintés funkcionális anatómiai jellegű és a később bemutatásra kerülő mesterséges modellekben implementált megoldásokkal való összevetést szolgája.



2.4. ábra. Az emberi hallórendszer sematikus felépítése.(O'Saughessy, Douglas „Speech Communications: Human and Machine" Copyrigth by Prentice Hall, Upper Saddle River, New Jersey)

Az ember hallórendszerének vizsgálatát az emberi test, a fej, illetve a fülkagyló morfológiai tulajdonságaival kell kezdeni, mivel ezen képletek akusztikus környezetre gyakorolt hatása, az egyes frekvencia-komponensek különböző arányú csillapítása révén fontos szerepet játszik a hangforrás irányának meghatározásában [12]. A hang, pontosabban a mechanikai rezgés a hallójáraton keresztül éri el a dobhártyát, ahonnan a hallócsontok közvetítése révén az ovális ablakon keresztül jut a csigába, pontosabban a cochlea vestibuláris csatornájában levő folyadékba [13]. A hanghullámok megrezgette folyadék mozgása átadódik a Corti-szervben levő alaphártyára, melynek mechanikai mozgása kitéríti a szőrsejteken levő stereociliumokat, melyek a mechano-szenzitív kation csatornák kinyitása révén a szőrsejteket depolarizálja. A belső szőrsejtek depolarizációja fázistartó elektromos impulzusok sorozatát eredményezi a hallóidegben. A cochlea mechanikai tulajdonságainak eredményeként [14] az alaphártya adott szakasza, a bejövő rezgés megfelelő frekvencia-komponensének energiájával arányosan mozdul ki, ezért a belső szőrsejtek, illetve a hallóideg a beérkező jel spektrális dekompozícióját továbbítja.

A hallásunk széles dinamika tartományának biztosítását a Corti-szervben elhelyezkedő külső szőrsejtek által, mechanikai visszacsatolás révén létrejövő adaptív erősítés biztosítja. A külső szőrsejtek esetén a depolarizáció eredménye ugyancsak kálium beáramlás, azonban ebben az esetben a beáramló kálium aktiválja a külső szőrsejt belső sejtvázát (citoskeleton), ami a szőrsejt alakváltozásához, elektromechanikus transzdukcióhoz vezet. A külső szőrsejtek stereociliumai elérik a basiláris membrán felett elhelyekezkedő membrána tectoria-t, aminek következménye a környező struktúrák,



2.5. ábra. A cochlea felépítése.(forrás: http://www.sfu.ca/sonic-studio/handbook/Cochlea.html)

elsősorban a membrana tectoria elmozdulása. A külső szőrsejt kontrakciója a passzív vándorló hullám [14] maximumának helyén növeli a kitérés nagyságát és kiemeli a szakaszt, így téve élesebbé a hangolást. A fenti bidirekcionális transzdukció érzékenyíti a szomszédos belső szőrsejtet, ennek a stereociliumai térnek ki nagyobb mértékben. A belső szőrsejtek ingerküszöbe 50-60dB-el magasabb mint a külső szőrsejtekét, ezért a passzív vándorló hullám amplitúdója a szoksásos hangnyomás szinteken (pl. beszéd) nem éri el a belső szőrsejtek ingerküszöbét, tehát fiziológiás hallásunk a külső szőrsejtek erősítő hatásának következménye.

A mesterséges rendszerek viszonylagosan gyenge teljesítményének megértése érdekében fontos megjegyezni, hogy mind a belső, mind a külső szőrsejteken efferens neuronok is végződnek, melyek a hallórendszer felsőbb régióiból (superior olivo complex) erednek, azaz a hallási afferentációt a központi idegrendszer receptor szinten is képes gátolni a transzmitter felszabadulás csökkentése révén. Annak eldöntésére, hogy ez az efferens rendszer az egész hallástartománynak az érzékenységét állítja-e, vagy képes szelektív gátlással az egyes frekvenciatartományokat, felsőbb kognitív folyamatoktól irányítva a környező frekvenciákból kiemelni, még nincs elegendő adat [13].

A térbeli hallás, illetve a jelek térbeli szegregációjának vizsgálatakor fontos megemlíteni a medialis olivo-cochlearis köteg neuronjait. Ezek a neuronok az ellenkező oldali cochleában végződnek és ugyancsak az afferens aktivitást csökkentik, a külső szőrsejtek bidirekcionális aktivitásának gátlása révén. A köteg funkciója egyenlőre nem tisztázott. Elképzelhető, hogy szerepe pusztán protektív, azaz a cochlea túl erős hangoktól való védelmét szolgálja, ugyanakkor az sem kizárható, - mivel az aktivitás frekvencia szelektíven változtatható - hogy funkciója, adott irányban levő forrás jelének kiemelése.

A térbeli hallásélmény kérdése kapcsán érdemes kiemelni a medialis superior oliva magot, melyben a hang érkezési irányára szelektíven tüzelő neuroncsoportokat találtak [15]. Feltételezések szerint ezen mag kulcsszerepet játszik a források térbeli elhelyezkedésének érzékelésében.

Összefoglalásként elmondható, hogy a hallórendszer mélyreható anatómiai vizsgalatának eredményeként tudjuk, hogy a felszálló ingerületek mellett, csaknem minden szint az alatta levőkhöz

leszálló ingerületeket is küld, azaz az összeköttetések reciprok jellegűek, tehát a rendszer egészét tekintve valószínűleg fontos a felsőbb régiók irányító szerepe. Az auditív perifériához csatlakozó komplex neuronális rendszer anatómiai felépítése, az egyes neuronok átcsatolódási helye aránylag jól ismert, de a hangészlelés, a hallásélmény kialakulását eredményező analízis módjáról a többi érzékszerv esetéhez viszonyítva kevesebb ismeretünk van.

2.3. A hallórendszer mesterséges modelljei

Bregman könyvének [6] megjelenése egy sor, a *hallási jelenet elemzés* filozófiáját alkalmazó mesterséges modell létrejöttét inspirálta[11] [7–10, 16–38]. Ezen munkák, mind az implementáció módját, mind a megvalósított csoportosítási szabályokat tekintve igen változatos képet mutatnak.

2.3.1. A frekvencia-idő reprezentáció

A hallórendszer csoportosítási mechanizmusait implementáló mesterséges rendszerek a hang reprezentációját tekintve két módszert alkalmaznak. Az egyik kategóriába a perifériás hallórendszer funkcionális modelljének kimenetét hasznosító eljárások tartoznak, míg a másik kategóriába az előbbihez képest bizonyos információ veszteséget okozó, ám kisebb számítási igényű módszerek sorolhatóak.

Szegmensenkénti Fourier transzformáció

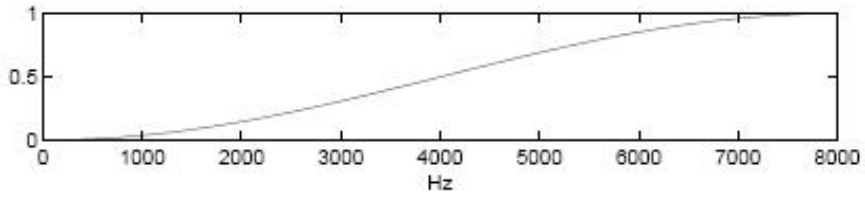
Egyes implementációkban [9,21,26–28] az akusztikus jelet szegmensenként frekvencia-komponenseire bontják, majd az egyes komponensek abszolut értékéből egy kétdimenziós energia térképet, úgynevezett *spektrogram*ot készítenek (a részleteket lásd a 4.1 fejezetben). A csoportosítási szabályokat az így létrehozott frekvencia-idő térképen kirajzolódó alakzatokként értelmezik. A módszer jellemzője, hogy a Fourier transzformációnak [39] köszönhetően hatékonyan számítható, azonban nem őrzi meg az egyes frekvencia-komponensek fázisát, mely binaurális feldolgozás esetén fontos információt hordoz a hangforrás helyére vonatkozóan. A módszer valamelyest módosított verziója a Q-sinus transzformáció [10], ahol a frekvencia-komponensek lokális maximumai adják a bináris frekvencia-idő reprezentációt. A spektrogram frekvencia tengelye - a cochlea frekvenciafelbontását megközelítendő - logaritmikusan választott, mivel így érhető el a gyakorlati szempontból legtöbb információt hordozó frekvencia-komponensek legrészletesebb ábrázolása.

A perifériás hallórendszer modellje

A hallórendszer funkcionális modelljével létrehozott frekvencia-idő reprezentációt jónéhány mesterséges modell alkalmazza [7, 19, 31, 35, 37]. Ennek oka azon feltevés, hogy a biológiai rendszer hatékonyságát nagyban meghatározza a kezdeti transzformáció információ-tömörítő és -kiemelő hatása.

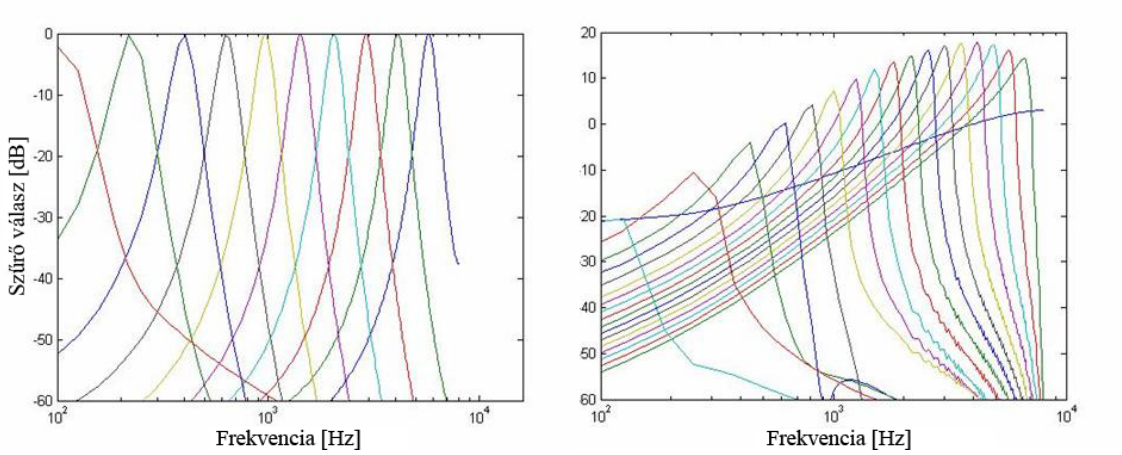
A perifériális hallórendszer analógiájára létrehozott számítógépes rendszerekben általánosan elhanyagolt a fülkagyló okozta irányszelektivitás. A modellek jellemzően a külső-, illetve a középfül szelektív frekvencia erősítő hatását visszaadó szűrőtömbbel kezdődnek (lásd 2.6. ábra).

[11] Computational Auditory Scene Analysis (CASA)



2.6. ábra. A középfül felül-áteresztő hatását modellező szűrő [19].

A következő lépcső a cochlea frekvencia-felbontó képességét modellező egy, az emberi hallás kritikus csatornái[12] alapján [40] paraméterezett szűrőtömb [41–43] (lásd 2.7. ábra). Ezek a modellek egy kivétellel [43] elhanyagolják a külső szőrsejtek bidirekcionális transzdukciós mechanizmusa révén jelentkező adaptív erősítési funkciót, jóllehet ez utóbbi munkában is meglehetősen heurisztikus módon kerül implementálásra, ezért csak a dinamika tartomány kiterjesztésének eszközeként használható. Érdemes megjegyezni, hogy létezik egy, a csigában levő folyadék dinamikájának leírása révén működő cochlea modell is [44], ez azonban a megoldás számítási igénye miatt napjainkban csupán elvi jelentőségű, annak ellenére, hogy az ezen modellel készített hangok frekvenciaképén találhatóak a későbbi feldolgozási lépések által hasznosítható finom időbeni minták, melyeket viszonylagos sikerrel használtak magánhangzó csoportok klasszifikációjára [45]. Ezen minták biológiai relevanciája azonban egyenlőre nem tisztázott, ezért elterjedt nézet, hogy a pszichoakusztikai megfigyelések alapján paraméterezett [46] szűrőtömb alapú megoldások kielégítő funkcionális modelljét adják a perifériás hallórendszernek.



2.7. ábra. A cochlea frekvenciafelbontását modellező szűrőtömbök. bal [41] ; jobb [42]

A közeltmúltban közölték a külső szőrsejtek matematikai leirását is [47], azonban ezek a dinamika tartomány alkalmas kiterjesztésén túl, a magasabb hierarchiájú irányító folyamatok ismere-

[12]Frekvencia sávok, melyeken belül egy adott frekvencia-komponens megléte érzékelhetetlenné teszi a sávon belüli egyéb összetevőket.

ténak hiánya miatt vélhetőleg egyelőre nem hoznak forradalmi áttörést a perifériás hallórendszer modellezésében.

Széles körben alkalmazottak a belső szőrsejtek mechanikus mozgást idegi impulzusokká alakító viselkedésének mintájára működő matematikai modellek [48, 49]. A szűrőtömbök kimenetén a negatív félhullámok levágása[13] révén szimulálják a szőrsejtek stereociliumainak egyirányú kitérésre érzékeny kation csatorna vezérlését, valamint az 5 kHz alatti hangok esetén a fázis-információt megőrző idegi impulzusválaszt. Az ion csatornák, valamint az intracelluláris folyadékban jelenlevő ionok számának matematikai modellbe való integrálásával megfigyelhető a hallóideg sztochasztikus tulajdonságokkal is rendelkező idegi impulzusválasza. A *hallási jelenet elemzés* filozófiáját követő, a perifériás hallórendszer modelljét alkalmazó megoldások mindegyike a hallóideg idegi impulzusait használja bemenetként.

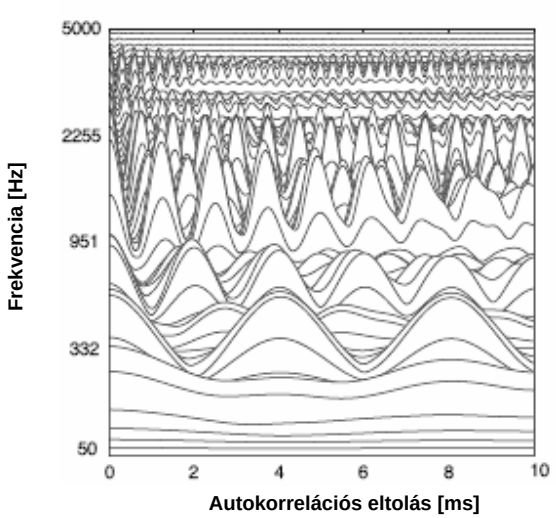
2.3.2. A primitív pszichoakusztikus csoportosítási szabályok implementációja

A mesterséges modellekben implementált pszichoakusztikai szabályok kiválasztása elsősorban a tervezett felhasználás követelményeitől függ. A legtöbb kísérletet a beszéd alapfrekvenciájának[14] azonosítására tették [7, 8, 10, 16, 17, 19, 21, 27, 29, 31, 32, 35, 36], mely a harmonikus tartalom és a közös frekvencia moduláció 2.1.1. fejezetben említett primitív csoportosítási szabályok együttes megvalósításával azonosítható. A kísérletek eredményeként sikerült legfeljebb két, egyidőben beszélő ember hangját szétválasztani, bár - mint Parison megjegyzi [31] - a szétválasztás hibája drámaian megnőtt a hangok alapfrekvenciáinak közelsége esetén. Az 2.3.1. fejezetben említett Fourier transzformációt használó megoldások a harmonikus kapcsolatot a frekvencia-idő képen végzett egyfajta alakzat-felismerési problémaként oldják meg [50,51]. A módszer hátránya, hogy a logaritmikus frekvenciafelbontás miatt a felső harmonikusok detekciójára alkalmatlan, mivel az egyes komponensek már nem megkülönböztethetőek, annak ellenére, hogy a hallórendszer kitűnően boldogul ezekben az esetekben is [52].

A hallórendszer számítógépes modelljét alkalmazó munkákban a Meddis által publikáltakhoz [53,54] hasonlóan járnak el. Az eljárások lényege, annak kihasználása, hogy a szűrőtömb kimenetén megjelenő félhullámok, illetve az ebből képzett impluzusok távolsága harmonikus hangok esetén egymásnak egész számú többszörösei. Klasszikus a Meddis által bevezetett megoldás, ahol a szűrötömb kimenetén csatornánként auto-korrelációt számolnak. Ez utóbbit *korrelogram*nak nevezik (2.8 ábra.). A korrelogramon a csúcsok időtengelyen való illeszkedése alapján meghatározható a harmonikus komplex alapfrekvenciája, illetve kiválaszthatók a komplexet alkotó felharmonikusok. Az eljárás fontos sajátossága, hogy az emberi hallgatókhoz hasonlóan, az alapharmonikus hiánya esetén is azonosítja az alapfrekvenciát. A módszer biológiai relevanciájának megkérdőjelezése okán született egy alternatív megoldás [55] az alapharmonikus detekciójára, jóllehet ennek biológiai relevanciája hasonlóan vitatott [56].

A jelek szinkron kezdetének/végének azonosítása viszonylag könnyen megvalósítható a spektrogramot bemenetként alkalmazó módszerek esetén [9,26–28], mivel a feladat a különböző frekvenciákon bizonyos szint felett megjelenő/eltűnő energia komponensek azonosításaként megvalósítható. A Sheffield Egyetemen készített dolgozatokban [8,16,19] a hallóideg impulzusait felhasználva valósítják meg az időben szinkron kezdődő komponensek kiválasztását, azonban nem önálló algoritmusként,

[13] half-wave rectification
[14] pitch



2.8. ábra. A korrelogram, egy úr és és hölgy beszédének elegyéből. A 3.8 ms-nál levő csúcsok a hölgy hangjának 253Hz alapharmonikusát, míg a 8.1ms-nál levő csúcs az úr hangjának 123Hz-es alapharmonikusának következménye [20].

hanem az alapharmonikus detekciójára használt korrelogram azonos idejű csúcsainak azonosításaként, a harmonikus tartalomra utaló periodicitás kizárásának segítségével.

A folytonosság csoportosítási szabály jellemzően a spektrogramot bemenetként alkalmazó munkákban kapott kitüntetett figyelmet [9,26,28]. A probléma ebben az esetben jól vizualizálható (lásd 4.22. ábra), adott küszöb feletti energia komponensek bizonyos méretű szakadásaként. A megvalósítás bonyolultsága és sikeressége elsősorban a jellemző frekvencia-komponensek reprezentációjának kérdése, ebből a szempontból pedig a multiágens szemléletet alkalmazó dolgozat tekinthető a legsikeresebbnek [28].

A zajszerű, a fenti módszerekkel nem azonosítható zárhangok, zöngétlen mássalhangzók komponenseinek csoportosítására Ellis [10] vezette be a közelség néven ismert szabályt, illetve hasonló motiváció szülte Wang munkáját [34]. Az implementációk hátránya az alkalmazott hardver környezetben való jelentős számításigény.

2.3.3. A hallórendszer felsőbb régióinak modelljei

Adavezérelt rendszer

Az adatvezérelt rendszerekben[15] [7,19,26] a fentebb bemutatott csoportosítási szabályok kimenetét közvetlenül igyekeznek a kívánt alkalmazás megvalósítására felhasználni. A létrehozott csoportokat pusztán a bemeneti adatokat figyelembe véve alakítják ki mellőzve minden szemantikai, vagy a

[15] bottom-up vagy data driven

csoportok egymásra való hatására vonatkozó vizsgálatot. A sikeresen megvalósított alkalmazások általános jellemzője, hogy csak szigorú, a hétköznapi felhasználási lehetőségeket kizáró feltételek megléte esetén, illetve csak speciális feladatok megoldására alkalmazhatók.

Állapotfüggő modell

Wientraub [37] még Bregman könyvének megjelenése előtt felismerte a pszichoakusztikai csoportosítási algoritmusokban rejlő lehetőségeket, és elsőként hozott létre a *hallási jelenet elemzés* metodikáját alkalmazó, néhány hang szeparációjának megoldására alkalmas számítógépes modellt. A pusztán adatvezérelt mechanizmusok gyengeségét felismerve néhány jelféléből, és azok átmeneti valószínűségeiből korlátozott világ-modellt alkotott, melynek aktuális állapota alapján valósította meg a beérkező hangok előre definiált jelekre való bontását. Munkájával bizonyította a primitív csoportosítási szabályok magasabb hierarchiájú szempontok szerinti kiválasztásának szükségességét.

Blackboard rendszerek

A blackboard rendszer a mesterséges intelligencia kutatásban használt architektúra. A megközelítés *hallási jelenet elemzés* metodikájában, a primitív csoportosítási algoritmusok (szoftver ágensek) eredményeinek valamely globális hipotézisbe való illeszkedése szerinti iteratív dekompozíciót jelent [8, 22]. A rendszerek teljesítményét erősen befolyásolja a rendelkezésre álló szoftver ágensek teljesítménye, illetve a beérkező jelek komplexitása.

Neurális hálózat alapú megoldások

Mint minden komplex alakzat-felismerési probléma megoldására, úgy a *hallási jelenet elemzés* problémájának megoldásában is megkísérelték a mesterséges neurális hálózatok képességeinek kiaknázását. A képfeldolgozással kapcsolatos problémák meglétén túl további nehézséget jelent az egyes jelek egymáshoz viszonyított, adott csoporton belül is megjelenő időbeni varianciája. Mindezen nehézségek ellenére születtek megoldások, melyek a jelek időbeni változására invariáns transzformációt - az auto-korrelációs függvényt - felhasználva sikeresen alkalmazták a neurális hálózatokat [35]. A probléma megoldásának egyik érdekes példája [18], ahol a szerzők a cochleáris szűrőtömb kimenetét egy kaotikus neurális oszcillátorokból álló hálózat gerjesztésére használnak. Mivel az azonos időben kezdődő komponensek szinkron gerjesztést biztosítanak, az érintett frekvenciákhoz tartozó elemek korrelált oszcillációba kezdenek.

Elvárásvezérelt megközelítés

A blackboard és az állapotfüggő modell elemeinek ötvözetéből kialakított rendszer [10]. A blackboard módszer globális optimum keresési stratégiájához hasonlóan iteratívan értékeli ki a primitív csoportosítási szabályokkal képzett objektumokat, ugyanakkor az állapotfüggő modellhez hasonlóan véges elemszámú (zaj felhő, zöngés hangok, tranziensek) lehetséges jelből összeállónak tekinti a bemenetet, ami alapján azután a dekompozíció történik.

2.3.4. Binaurális információk integrációja

A binaurális információk felhasználásának jelentősége abban áll, hogy a csoportosítási szabályok által létrehozott különálló hangobjektumokat viszonylag könnyen lehet a kibocsátó forrás irányára

vonatkozó információk alapján összetartozó érzékelési folyamokba rendezni, azaz például egyes fonémákból szavakat alkotni. Ennek egyik példája Nakatani munkája [29], amelyben a harmonikus komponensek érkezési irány szerinti elkülönítését valósítja meg. Itt érdemes megemlíteni azokat a munkákat, melyek az idő[16], illetve az intenzitás[17] különbséget felhasználva igyekeztek a forrás helyére vonatkozó információkat szerezni. Ezen kutatások elsődlegesen a cochlea protézisekben, illetve hallókészülékekben való felhasználást célozták [57, 58], ugyanis az érdektelen környezeti zajok felerősítése erősen rontja a nagyothallók életminőségét. Mivel a citált munkák elsősorban a *hallási jelenet elemzés* problémáinak megoldására koncentráltak, nem foglalkoztak a visszhang jelentőségével, ami - mint az 5. fejezetben részletezem - alapvető fontosságú a gyakorlatban használható forrás-lokalizáló algoritmusok esetében. Fontos előrelépés volna a binaurális információk, a magasabb hierarchiájú kontroll funkciókhoz hasonló, a primitív csoportosítási szabályok eredményének iteratív kiértékelésében való alkalmazása.

2.3.5. A számítási modellek teljesítményének összevetése

A különböző implementációk összehasonlítása nehéz feladat, mivel némi variabilitás tapasztalható az egyes szabályok gyakorlati körülmények között való értelmezésének módjában. A szabályok implementációjakor a szerzők egy adott feladat megoldására koncentráltak, ezért lehetetlen ugyanazon szabály különböző implementációinak különböző körülmények közötti különböző feladatok megoldásában nyújtott teljesítménye alapján összehasonlítást végezni, különösen, hogy összemérhető kvantitatív adatok nem publikáltak. Mint arra Ellis rámutat [10], a *hallási jelenet elemzés* motiválta számítógépes megoldások száma elérte azt a szintet, amikor szükség lenne az egyes rendszerek egymáshoz viszonyított teljesítményének meghatározására. Az igény tehát megvan egy szabványos teszt jelkészlet összeállítására, mint az a beszédfelismeréssel kapcsolatos kutatások estén jól működő gyakorlat, azonban egyelőre ilyen nem létezik. Ennek oka vélhetően az, hogy az implementációk rendkívül sokrétűek, bonyolultak, ezért komoly energiát igényel a modellek ismételt implementációja, ami lehetővé tenné az egységes teszt környezetben való validációt. Összehasonlításra ebben a dolgozatban sem vállalkozom. A rendszerek teljesítményét illetően azonban annyi mindenképpen elmondható, hogy az implementált megoldások általában elvi jelentőségűek, céljuk egy adott funkció *hallási jelenet elemzés* módszereivel történő megvalósítása. A gyakorlati felhasználás szempontjából lényeges paraméterek, mint a számítás igény vagy futásidő a tárgyalt dolgozatok egyikében sem vizsgált.

2.4. Konklúzió

A *hallási jelenet elemzés* bemutatott megvalósításai igazolják, hogy a hallórendszerről jelenleg rendelkezésünkre álló tudás alapján készíthetőek, javarészt csak speciális körülmények között és/vagy csak bizonyos feladatok megoldására alkalmas algoritmusok. Az algoritmusok célja jellemzően a több forrás jeléből álló elegy összetevőkre bontása, majd az egyes hangesemények felismerése, beszéd esetén az alapharmónikus detekciója, illetve elterjedt az azonosított komponensek forrás-helyének meghatározása. Ez utóbbi alkalmazásra mutatok példát a 4.4. fejezetben.
A létező módszerek mindennapi életben való alkalmazásának hátránya, hogy a hallórendszer működését alapvetően befolyásoló magasabb hierarchiájú folyamatok egyenlőre ismeretlenek. További

[16] Interaural Time Difference (ITD)
[17] Interaural Level Difference(ILD)

probléma, hogy az azonosított szabályok mesterséges modelljeinek számításigénye napjainkban kizárja a valós idejű alkalmazás lehetőségét, valamint lehetetlenné teszi a tanult, illetve kontextusfüggő vezérlés alapján történő primitív csoportosítási szabályok eredménye alapján végzett iteratív dekompozicót. A vázolt nehézségek egy részére megoldást nyújthat az alternatív számítási paradigmák felhasználásának lehetősége, melyről a következő fejezetekben értekezem.

3. FEJEZET

A Celluláris Hullámszámítás

A digitális számítógépek teljesítményének rohamos fejlődését lehetővé tevő "scaling down" jelenség a közeljövőben már nem biztosít lehetőséget a digitális processzorok órajelének, illetve az egységnyi felületen elhelyezhető tranzisztorok számának növelésére, holott a szenzorok, szenzortömbök szolgáltatta több dimenziós adatfolyamok valós idejű analízise - ésszerű korlátok között - a napjainkban rendelkezésre álló számítási teljesítménnyel nem lehetséges. A probléma felismerését követően számos alternatív számítási paradigma alkalmazására történt kísérlet.

Ezek talán legsikeresebbike a Celluláris Neurális Hálózatok (CNN) elmélete, melynek alapjait [59] és lehetséges alkalmazási területeit [60] ismertető munkákat Leon O. Chua és L. Yang 1988-ban publikálta. A CNN hálózatok képességeinek algoritmikus kihasználását a CNN univerzális gép [61] (CNN-UM) megszületése tette lehetővé. A CNN-UM-mel a celluláris rácson terjedő hullámok számítási potenciálját sikerült algoritmikus problémák megoldásában kamatoztatni. Az elmúlt, mintegy 20 év kutatómunkájának eredményeként számos, a hagyományos logikai, illetve szekvenciális algoritmus szervezéssel nehezen megoldható problémára sikerült rendkívül hatékony megoldást adni. Az eredmények, a klasszkus komplexitás fogalmak újraértelmezéséig vezettek [62]. Világossá vált, hogy az egyes feladatok komplexitása nem pusztán az őt megvalósító program aggregált számítási igényétől, a számítás végrehajtása közben disszipált teljesítménytől, vagy a programot megvalósító szilikon felület nagyságától függ. A megvalósítandó algoritmusok komplexitása csak az implementáció alapját adó architektúra figyelembevételével értelmezhető [62]. A kétdimenziós celluláris rácson lezajló tranziens folyamatok számítási potenciálja elsősorban képfolyamok, azaz több dimenziós, tér-időbeli problémák megoldásában aknázható ki.

A 2. fejezetben, az emberi hallás pszichoakusztikus megfigyelésekkel azonosított funkcióit és az azokat megvalósító modelleket mutattam be. A modellek mindegyike különböző frekvencia-idő reprezentációkon, azaz képfolyamokon végzett műveletek révén oldja meg a kívánt feladatot. A probléma megoldására a celluláris hullámszámítógép alkalmas lehet, ezért jelen fejezet hullámszámítással kapcsolatos alapelveket, valamint az algorimusok megvalósításával kapcsolatos praktikus szempontokat taglalja.

3.1. A Celluláris Neurális Hálózat

A CNN alkalmas feladatokban nyújtott megdöbbentő számítási teljesítményének kulcsa a szabályos rácsban elhelyezett processzorok párhuzamos számítási teljesítménye. CNN hálózatról beszélünk

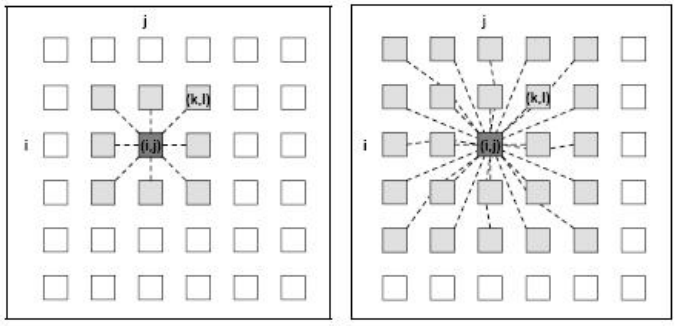
ha:

- A processzorok valamilyen szabályos geometriai struktúra rácspontjaiban helyezkednek el és csak egy meghatározott r sugarú véges környezettel vannak közvetlen kapcsolatban.
- Az időváltozó (t) lehet folytonos vagy diszkrét.
- Minden állapotváltozó (x) folytonos értékű.
- A programot, illetve ennek elemi utasításait a cellák bemenete (u), azok állapota (x), a kimenetek (y), illetve az szomszédos cellák bemenetei és állapotainak úgynevezett templatekkel megadott súlytényezői határozzák meg.

Az általános CNN architektúra egy MxN-es két-dimenziós négyzetrács rácspontjaiban elhelyezett processzorokból, cellákból álló tömb, ahol $C(i,j)$ a hálózat i-edik sorának j-edik oszlopában lévő cellát jelöli. $(i = 1\ldots, M;\ j = 1, \ldots, N)$. A $C(i,j)$ cella hatókörének területe $S_r(i,j)$, ami a szomszédos cellák olyan halmazát jelöli, amelyek megfelelnek a következő feltételnek:

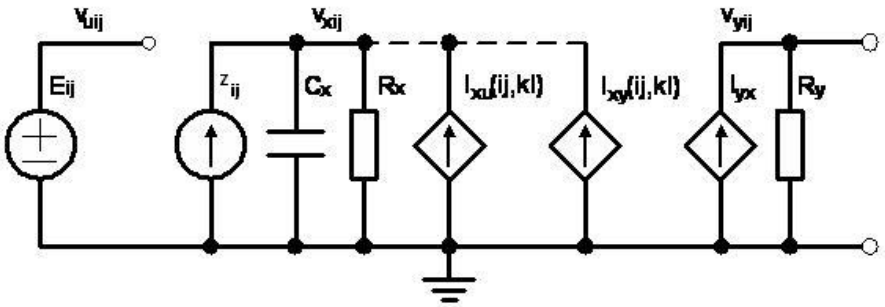
$$S_r(i,j) = \left\{ C(k,l) \left| \max_{1\le k\le M, 1\le l\le N} \{|k-i|, |l-j|\} \le r \right. \right\} \tag{3.1}$$

A 3.1 ábrán egy tetszőleges $C(i,j)$ cella $r = 1$ és $r = 2$ környezete látható.



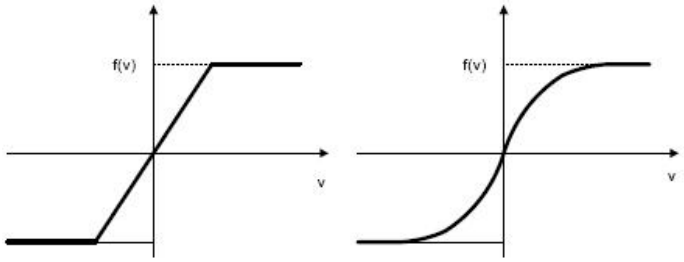
3.1. ábra. A CNN cella összeköttetési hálózata $r = 1$ (bal) illetve $r = 2$ (jobb) esetben.

A cellákat tömbön belüli elhelyezkedésük alapján két csoportra osztjuk, a belső cellák illetve a határoló cellák csoportjára. A $C(i,j)$ cellát belső cellának nevezzük, ha minden cella létezik az r sugarú környezetén belül, egyébként határoló celláról beszélünk. Az elsőrendű $C(i,j)$ CNN cella áramköri modellje a 3.2 ábrán látható. A v_{xij} csomóponti feszültség a $C(i,j)$ cella állapota, míg v_{uij} ugyanezen cella bemenete, v_{yij} pedig a kimenete. Látható, hogy minden cella tartalmaz egy független feszültségforrást (E_{ij}), egy független áramforrást (z_{xij}), egy lineáris kondenzátort (C), továbbá két lineáris ellenállást (R_x és R_y). Amennyiben a szomszédos cellák száma m, akkor tartalmaz még legfeljebb $2m$ lineáris feszültségvezérelt áramforrást, amelyek a szomszédos cellákhoz csatololtak a vezérlő feszültségeken keresztül, ahol v_{ukl} a szomszédos cellák vezérlő bemeneti feszültségei, v_{ykl} pedig a szomszédos cellák kimeneti visszacsatoló feszültségei. Bevezetve az $I_{xy}(ij;kl) = A(ij;kl) \cdot v_{ykl}$ és $I_{xu}(ij;kl) = B(ij;kl) \cdot v_{ukl}$ jelölést minden $C(i,j)$ és $N_r(k,l)$-re, az $A(ij;kl)$ a visszacsatoló (feedback), míg $B(ij;kl)$ az előrecsatoló (control) *template*. Az egyetlen nemlineáris elem minden



3.2. ábra. Az elsőrendű CNN cella áramköri modellje.

cellánál a szigmoid karakterisztikájú feszültségvezérelt áramforrás $\mathrm{I_{yx}} = \frac{1}{R_y}\mathrm{f(v_{xij})}$, amelynek két típusa látható a 3.3 ábrán. A 3.2. ábrán látható cellákból felépülő A és B template-ekkel megadott



3.3. ábra. Nemlineáris átviteli függvények. Szakaszonként lineáris (bal) és folytonos szigmoid karakterisztika (jobb).

rendszer dinamikáját az alábbi differenciálegyenlet-rendszer jellemzi:

$$C_x \frac{dv_{x_{ij}}(t)}{dt} = -\frac{1}{R_x} v_{x_{ij}}(t) + \sum_{C(k,l)\in S_r(i,j)} A_{ij;kl} v_{y_{kl}}(t) + \sum_{C(k,l)\in S_r(i,j)} B_{ij;kl} v_{u_{kl}}(t) + z_{ij} \qquad (3.2)$$

ahol $z_{i,j}$ az adott cellához tartozó úgynevezett *bias* áram értéke. A hálózat belső állapota a nemlineáris elemen keresztül határozza meg a cella kimenetét, ami - az elterjedtebb -, szakaszonkénti lineáris függvény esetében formálisan az alábbi alakban írható:

$$v_{y_{ij}}(t) = f(v_{x_{ij}}(t)) = \frac{1}{2}(|v_{x_{ij}}(t)+1| - |v_{x_{ij}}(t)-1|), i = 1...M\,; j = 1...N, \qquad (3.3)$$

Amennyiben $A_{ij;kl}$, $B_{ij;kl}$ értékei helyfüggetlenek, vagyis nem függnek i és j értékeitől, térinvariáns template-ekről beszélünk. A cellánként megadott *bias* áram a bias térkép[1]. Általános esetben az bias áram értéke helyfüggetlen ($z_{ij} = z$). A cella dinamikáját az RC tag valamint az A, B és z template-ek határozzák meg. A kimenet alakulására hatással van a bemenet (v_{uij}) és a kezdeti

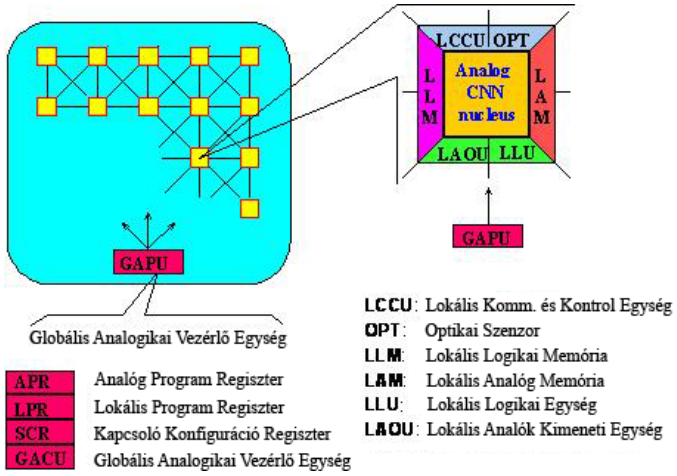
[1] bias map

állapot ($v_{xij}(0)$), ezért a hálózat tranziensének lezajlása, azaz a CNN művelet elvégézése előtt ezen paraméterek inicializálásáról gondoskodni kell.

Az eddig leírt CNN hálózatban a templatek konstans értékek, így az állapot, a kimenet és a bemenet lineáris kombinációja. A hálózat viselkedésére nemlineáris, időfüggő, illetve helyfüggő template érték esetére is ismertek tételek, azonban ezek tárgyalásától a hardver implementációk hiánya miatt eltekintek.

3.2. CNN Univerzális Gép

Bár a CNN hálózat VLSI megvalósítása nagy számítási sebességet tesz lehetővé, az igazán széleskörű alkalmazást az algoritmikus programozhatóság és a nagyfokú rugalmasság megteremtése teszi lehetővé. A CNN Univerzális Gép (CNN-UM) egy analóg tárolt programú celluláris tömbszámítógép - tulajdonképpen egy analóg mikroprocesszor - amely lokális analóg és logikai memóriával, önálló operációs rendszerrel és programozási nyelvvel rendelkezik [61].



3.4. ábra. A CNN Univerzális Gép - globális architektúra [63].

Az univerzális chip felépítése a duális számítási paradigmán alapul, ami az analóg operációk logikai műveletekkel, lokális analóg memóriákkal és programozhatósággal való kombinációját jelenti. Az analóg és a logikai számítás ötvözetének elnevezésére az *analogikai számítás*, *analogikai algoritmus* kifejezést használjuk. A hibrid számításokkal ellentétben, ebben az esetben nincs szükség A/D és D/A átalakításra, az analóg értékeket nem kell digitálisan reprezentálni, minden jel és operátor vagy analóg, vagy logikai. A 3.4. ábrán látható a CNN univerzális gép felépítése, amely a központi vezérlő (GAPU) segítségével hangolja össze a rácsban elhelyezett processzáló elemek működését. A CNN cellákhoz tartozik egy-egy lokális analóg memóriaegység (LAM), néhány analóg memóriaelem (LAMi), a lokális logikai memóriaegység (LLM) és a lokális kommunikációs és kontroll egység (LCCU). Ez utóbbi biztosítja a kapcsolatot a központi globális analogikai vezérlő egységgel (GAPU). A cellában lévő lokális logikai egység (LLU), illetve lokális analóg kimeneti egység (LAOU) bemeneteit a LAM, illetve LLM memóriákból veszi. Az analóg memóriaelemek biztosítják a CNN hálózat iteratív alkalmazásának lehetőségét, hiszen adott bemenetre adott template-et

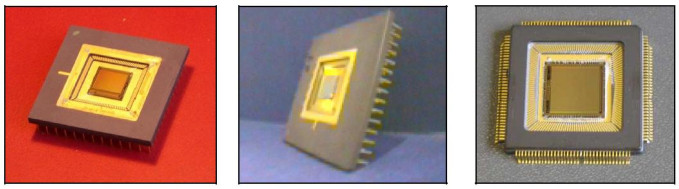
futtatva az eredmény lokálisan tárolható, majd ezeket felhasználva indítható a következő művelet. A központi analogikai vezérlő egység tartalmazza az utasításregisztereket, melyekben helyet kap az analóg program regiszter (APR) is, mely a template-eket tárolja. A logikai utasítások logikai programregiszterben (LPR) tárolódnak. A bináris kapcsoló konfiguráció regiszter (SCR) a cellában lévő kapcsolók állását kódolja, amivel a bemenetként, illetve kimenetként használatos LAM-ok, illetve LLM-ek címezhetők.

3.3. A CNN-UM hardver megvalósításai

A celluláris hullámszámításban rejlő lehetőségek kulcsa a nagy számítási teljesítmény, ezért fontos szempont a kidolgozott algoritmusok gyors futtatását lehetővé tevő hardver elemek megléte. Az elmúlt évtizedben a világ több CNN kutatással foglalkozó laboratóriumában készítettek a CNN paradigma alkalmazását lehetővé tevő chip-eket [64–73]. Kezdetben szilikon alapú, nem programozható, illetve részben programozható, majd a későbbiekben univerzális CNN chip-ek készültek, részben analóg VLSI, részben emulált digitális technikával. Az analóg chip-eket nagy számítási teljesítmény, ugyanakkor korlátozott pontosság, nagy zajérzékenység, míg a digitálisan emulált implementációkat kellő fokú precízitás, ám valamelyest csökkent számítási sebesség jellemzi.

Gyártás helye	Készítés időpontja	CNN tömb mérete	Cella típusa	Alkalmazott technológia [m]	Idő állandó (τ) [s]	Bement
Berkeley & Müchne	1993	12×12	DTCNN	2µ	300ns	analóg
Seville	1994	32×32	Full-range	1µ	-	bináris & optikai
Leuven	1995	20×20	Chua-Yang	0.7µ	4.8µs	analóg
Seville	1995	20×22	Full-range	0.8µ	400ns	bináris & optikai
Berkeley	1996	16×16	Chua-Yang	1µ	27ns	analóg
Helsinki	1997	48×48	Chua-Yang	0.5µ	50ns	bináris
Seville	1998	64×64	Full-range	0.5µ	250ns	analóg
Seville	2000	32×32	Full-range, complex cella	0.5µ	<100ns	analóg
Seville	2001, 2004	128×128	Full-range	0.35µ	250ns	analóg & optikai
Budapest	2004	n×40	Full-range	0.35µ	-	digitális

3.1. táblázat. A különböző CNN implementációk teljesítményének alakulása [74].



3.5. ábra. A CNN univerzális chip-ek. Balról: Seville 20x22 (1995); Ace4k (1998); Ace16k (2001) [74].

3.4. CNN algoritmus tervezése

A digitális számítógépek programozásánál megszokottól eltérő gondolkozást igényel a CNN paradigmára épülő analogikai algoritmusok tervezése. A digitális számítástechnikában a programot az aritmetikai és logikai műveletek szekvenciális sorozata adja, míg az analogikai programokat logikai és tér-időbeli analóg operációk kombinációi építik fel. Az analóg operáció, a celluláris rácson terjedő hullám viselkedése ugynevezett template-ekkel határozható meg egy adott feladatban (lásd a 3.4. egyenlet és a 3.6 ábrák).

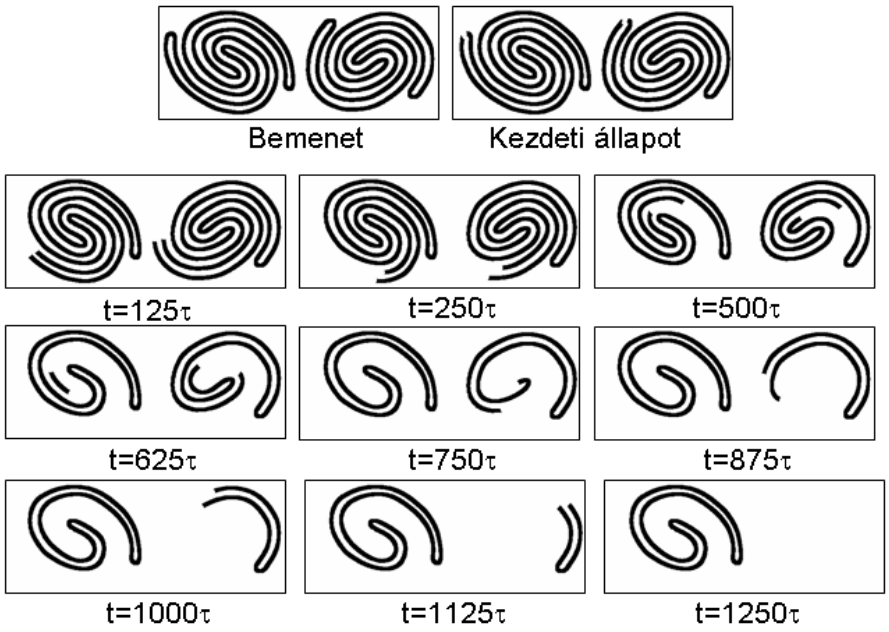
$$A = \begin{bmatrix} 0 & 0.5 & 0 \\ 0.5 & 3 & 0.5 \\ 0 & 0.5 & 0 \end{bmatrix} B = \begin{bmatrix} 0 & -0.5 & 0 \\ -0.5 & 3 & -0.5 \\ 0 & -0.5 & 0 \end{bmatrix} z = -4.5 \tag{3.4}$$

Azon a template-eket melyeknél az A mátrix a centrális mezőt kivéve csak nulla elemeket tartalmaz nem propagáló template-eknek nevezzük. Ezekben az esetekben a cella kimenetét csupán a szomszédos cellák értéke határozza meg. Propagáló template esetén az egyes cellák hatása az A mátrix előrecsatolása miatt az alkalmazott template szomszédsági körén kívül is kifejtheti hatását. A analóg program kiválasztásakor, azaz a template tervezésnél, illetve a meglevők kiválasztásakor fontos szempont a kívánt művelet zajjal szembeni robusztussága, mivel az egyes hardver elemek korlátozott pontossága lehetetlenné teheti a felhasználást.
A hatékony analogikai algoritmus-tervezés a következő szempontok figyelembevételét igényli:

- A párhuzamos feldolgozás javarészt lokális interakciók révén valósuljon meg.
- A közbenső eredmények lokálisan legyenek eltárolva.
- A döntések globális paraméterek (pl: minden képpont fehér) függvényében történjenek, mivel ezek detekciója egyszerűen és gyorsan megvalósítható.

Az analogikai algoritmusok fejlesztésére felhasználóbarát, a különböző paraméterek hatásának vizsgálatát megkönnyítő szoftver szimulátorok állnak rendelkezésre [76]. A szoftver szimulátorok az alkalmazott CNN hardver platformtól függetlenül, ugyanakkor annak speciális sajátosságait figyelembe véve teszik lehetővé a programfejesztést. A CNN chip-ek egy, a felhasználó számára transzparens hardver-szoftver interfészen keresztül érhetőek el, ami megkönnyíti a különböző szilicium implenetációkon futó algoritmusok platformfüggetlen fejlesztését. Az interfészek korai verziói PC-be ágyazható kiegészítő kártyák voltak, azonban ma telepről üzemeltethető, hálózaton keresztül konfigrálható önálló eszközök.

A rendszer különböző absztrakciós szinteken férhető hozzá. A legalsó szint a C++ függvényhívások alkalmazása, melyek adott feladatok CNN chip-en való elvégzését, vagy a CNN képfeldolgozó



3.6. ábra. Példa a celluláros rácson terjedő hullám számítási teljesítményét hatékonyak kiaknázó feladatra. A ??. ábrán látható *GlobalConnectivityDetection* [75] template-et felhasználva a eltüntethetőek azok a bemeneti objektumok, melyek a CNN rács kezdeti állapotában fehér pixelel jelöltek.

könyvtár [77] komplex rutinjainak futtatását eredményezi. A következő absztrakciós szint az úgynevezet AMC (Analogic Micro Code) kód, mely lehetővé teszi a CNN-UM chip-ek regiszter-szintű hozzáférését. Az Alpha nyelv az analogikai algoritmusok magasszintű programozási nyelve. Az algoritmus tervezés fontos eleme, egy már létező programkönyvtár [75], ami az egyes feladatok megoldására szolgáló template-eket és algoritmus példákat foglalja össze.

3.5. Az analogikai algoritmusok implementációs szempontjai

A hatalmas számítási teljesítménnyel bíró szilícium alapú CNN-UM chip-ek, az analóg VLSI gyártás sajátosságaiból fakadóan korlátozott pontosságúak, ami behatárolja a gyakorlatban futtatható template-ek számát. Az elmúlt évek kutatómunkájának hála, egy sor elmélet született a megtervezett template-ek robusztusságának vizsgálatára [78], ami tervezhetővé tette a bináris - csak fekete-fehér - bemenetetű és kimenetű template-ek analóg VLSI áramkörökön való futtatásának sikerességét. A chipekben fellépő elektromos és termikus zajok hatásának becslésére statisztikus modellek készültek [79, 80], illetve különböző dekompozíciós módszerek kerültek kidolgozásra [81], melynek eredményeként a viszonylag bonyolult visszacsatoló template-et is tartalmazó, ezért kevésbé robusz-

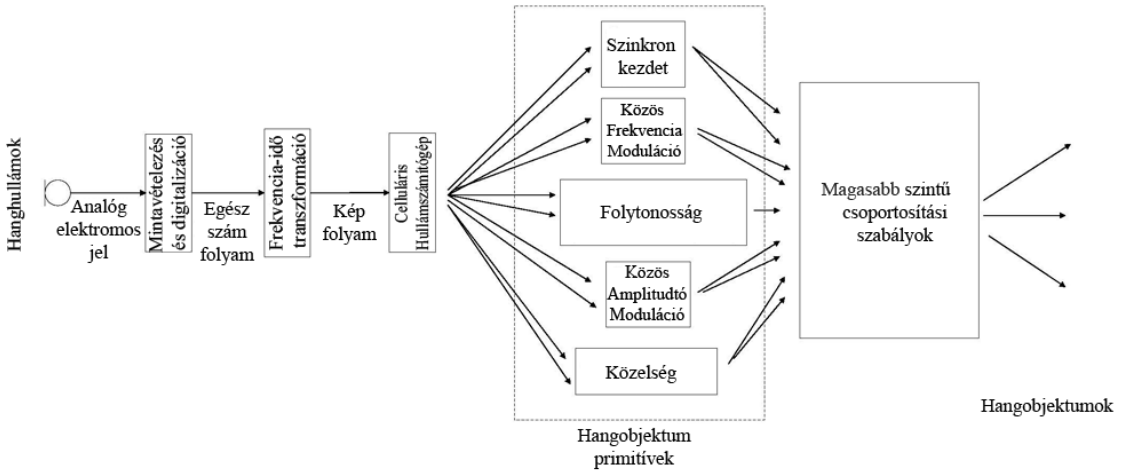
3.7. ábra. A különböző CNN chipek fogadását lehetővé tevő hardver platformok [74].

tus operációk is futtathatóak lettek bizonyos korlátozások mellett. Az egyik ilyen korlát volt - a sziliciumon való összeköttetések korlátozott számából fakadó - 3x3-as template-méretre vonatkozó megkötés, azonban ma már léteznek módszerek, melyek lehetővé teszik a nagyobb szomszédságú template-ek 3x3-as template szekvenciákkénti megadását [82]. Ugyancsak fontos eredmény, a chipeken definiálható helyfüggő bias térkép, ami lehetővé teszi a helyfüggő template-ek eredményének analóg tranziensek sorozataként való meghatározását.

A fentiek eredményeként, ma már jól ismertek az analogikai algoritmusok nagysebességű analóg VLSI áramkörökön való futtatásának feltételei. Bináris operációk 0.7 robusztussági szint felett biztosan alkalmazhatóak. Propagáló template-ek (az A template-nek van a középső elemet kivéve is nullától különböző értéke) alkalmazására egyelőre nem ismert általános módszer, a bináris template sorozatonként való futtatás - azaz nem propagáló bináris bemenetet és kimenetet produkáló template szekvenciával való megvalósítás - módját egyedileg kell meghatározni.

A HALLÁSI JELENET ELEMZÉS HULLÁMSZÁMÍTÁSI KERETRENDSZERE

A 2.3.2. fejezetben bemutatott *hallási jelenet elemzést* megvalósító alkalmazások példáján világossá vált, hogy a gyakorlatban alkalmazható, a feladathoz jobban illeszkedő architekturára van szükség. A 3. fejezetben a celluláris hullámszámítási paradigmát, mint lehetséges alternatív architektúrát tekintettem át, mivel az alkalmasan reprezentált problémák megoldásában napjaink szuperszámítógépeihez mérhető számítási teljesítményt nyújt. Az alábbiakban a *hallási jelenet elemzés* problémájának hullámszámítással történő megoldását részletezem. A fejezetben bemutatott módszerek és eljárások teljes egészében saját munkám eredményei, az esetleges hivatkozások a már meglévő tapasztalatokra utalások, illetve az alapként szolgáló eredményeket jelölik.



4.1. ábra. A *hallási jelenet elemzés* hullámszámítási keretrendszere.

A tárgyalt modell vázlata a 4.1. ábrán látható. A rendszer első eleme, hasonlóan a többi

számítási modellhez, a cochlea funkcionális analógiája szerint paraméterezett spektrális felbontás, mely az egyetlen mikrofon által rögzített analóg hangjelet, a digitalizálást követően fekete-fehér árnyalatú képfolyammá alakít. Ez a képfolyam a hullámszámítógép bemenete, míg a kimenet a primitív csoportosítási szabályok hullámszámítógépes implementációja révén létrejövő hangobjektum primitívek. Jelen dolgozat témája a hangobjektum primitívek létrehozásának mind hatékonyabb módja, ezért nem érintem a tanult, illetve kontextusfüggő folyamatok által vezérelt, a primitív csoportosítási szabályok eredménye alapján végzett iteratív dekompozicót. A 2.3.3. fejezet metodikája alapján a bemutatott modell az adatvezérelt rendszerek közé sorolható.

Az egyes funkciókat megvalósító algoritmusok lényeges pontjainak tárgyalása mellett minden esetben közlöm az analogikai algoritmusok leírására általánosan használt UMF diagramot [62]. A diagramokon belül szaggatott vonallal keretezve jelölöm a keretben levő elemek összességeként definiált, később a feltüntetett névvel hivatkozott szubrutinokat. A fejezetben definiált szubrutinok a szövegben **vastag** betükkel szedettek. Ugyancsak **vastag** betükkel írtam a CNN programkönyvtárban publikált súlymátrixok neveit, melyek fordítása esetén a lábjegyzetben tüntettem fel az eredeti angol elnevezést. Az UMF diagramok könnyebb követhetősége miatt a súlymátrixok angol elnevezéseit használtam. A szövegben és az UMF diagrammokban *dőlt* betükkel szedtem a CNN rács, köztes eredményekként elmentett pillanatnyi állapotait. Az algoritmusok szoftver szimulátoron tesztelt, platformfüggetlen AMC kódja a http://lab.analogic.sztaki.hu/awct oldalról letölthető.

4.1. A hang frekvencia-idő reprezentációja

Mint azt a 2.3.1. fejezetben említettem a *hallási jelenet elemzéssel* foglalkozó munkákban több módszer használt a beérkező hang frekvencia-idő reprezentációjának előállítására. Elődeim példáját követve magam is kidolgoztam egy transzformációt, melynek megalkotásánál az alkalmazott architektúra elvárásait tartottam szem előtt. Mivel célom egy minél gyorsabban működő rendszer létrehozása volt, a frekvencia-idő transzformáció kifejlesztésénél elsődleges szempont volt a hatékony kiszámíthatóság, ugyanakkor a cochlea funkcionális modelljét alapul véve igyekeztem biztosítani a feladathoz leginkább illeszkedő információ-reprezentációt. Ezen megfontolások alapján úgy döntöttem, hogy Fourier felbontást használok a szűrőtömb alkalmazása helyett, mivel ennek digitális számítógépen való kiszámítása lényegesen hatékonyabban elvégezhető.

Ennek hátránya a precíz időbeli információ elvesztése, mely a 2.3.2. fejezetben említettek alapján az alapharmonikus, illetve harmonikus tartalom emberi halláshoz hasonló robusztus detekciójának hiányában, illetve binaurális rendszerekben a forrás helyére vonatkozó információ elvesztése révén jelentkezik. Az alapharmónikus detekciója, a hangmagasság érzet kialakulásának magyarázata komplex kérdéskör, önálló kutatási terület, ezért úgy döntöttem nem kísérlem meg az itt bemutatandó egységes keretrendszerbe való integrálást, ezen csoportosítási szabály implementációjától eltekintek.

Jelölje tehát $s(i)$, $(i = 1...\infty)$ a digitalizált hang i. pillanatban felvett értékét. A digitalizált jelfolyamot bontsuk O mintával átfedő, W hosszúságú, Gauss fügvénnyel súlyozott ablakokra a következő egyenlet szerint:

$$w_p(k) = \mathrm{s(p} * \mathrm{(W - O) + k)} * \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(k-\mu)}{2\sigma^2}} \tag{4.1}$$

ahol p a szegmens sorszáma, $w_p(k)$ a p. szegmens k. időpillanatban felvett értéke, $k = 1...W$, valamint legyen $\sigma = W/4$, továbbá $\mu = W/2$.

Az ablakokon alkalmazott súlyozás a spektrális felbontás pontosságát hivatott növelni, mivel a Fourier felbontást az elmélettel ellentétben nem végtelen méretű szegmenseken számítom. Elterjedt a Hamming ablak haszálata, azonban ennek alkalmazása esetén a szűk frekvenciasávban bekövetkező hirtelen energia ugrások a spektrum egészére kiterjedő csúcsokat hoznak létre, ami esetemben kifejezetten hátrányos, ezért választottam a Gauss függvénnyel való súlyozást.

A Gauss függvénnyel súlyozott ablakok Fourier felbontását jelölje $S_p \in \mathbb{C}^{\frac{W}{2}}$, $W/2$ dimenziós komplex vektor. Mivel a frekvenciafelbontás ebben a vektorban lineáris, és mint a cochlea példáján látható, a logaritmikus skála jobban megfelel a követelményeknek, kompressziót végeztem az alábbiak szerint:

$$SL_p(j) = \sum_{i=N}^{M} \frac{|S_p(i)|}{M-N} \tag{4.2}$$

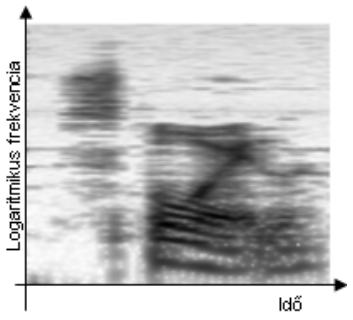
ahol $SL_p(j)$ a logaritmikus frekvenciafelbontású vektor j. komponense, $N = q^j$, $M = q^{j+1}$ és $j = 1...C$. C a kompressziót követően létrejövő csatornák száma, $|S_p(i)|$ az S_p komplex vektor i. komponensének abszolút értéke. Az SL_p valós értékű C dimenziós vektort jelöl. A logaritmikus kompressziót követően a j. csatorna a q^{j+1}-től a q^j-ig terjedő frekvencia tartományba tartozó jelek energiájának átlagát tartalmazza, ahol (q) a következő kifejezés alapján határozható meg:

$$\frac{\mathrm{W}}{2} = \sum_{i=1}^{C} q^i \tag{4.3}$$

A fenti eljárással nagyjából a cochleával, illetve a pszichoakusztikai megfigyelések alapján paraméterezett szűrötömbökkel azonos frekvencia felbontás érhető el (lásd 2.7. ábra), azzal a különbséggel, hogy ezúttal a szomszédos csatornák egymással nem átfedőek. A hang frekvencia-idő intenzitásképét az előzőekben kiszámolt valós értékű vektor komponenseinek logaritmusát véve kapom:

$$spect(p,k) = log(SL_p(k)) \tag{4.4}$$

Ez a hallóideg intenzitásfüggő tüzelési gyakoriságának, pontosabban a külső szőrsejtek révén létrejövő széles dinamikatartománynak a durva közelítése. A $spect(p,k)$ a p. ablak spectrális felbontásából képzett vektor k. komponensét jelöli, ami praktikusan a spektrogram (p,k) koordinátákon elhelyezkedő komponense. A transzformáció eredményét, [10]-hoz hasonlóan, férfi beszélő által kiejtett angol 'spoil' szó mintáján szemléltetem.



4.2. ábra. A 'spoil' szó spektrogramja.

A spektrogram felbontásának meghatározását két szempont figyelembevételével kell megtenni. A felbontásnak elég finomnak kell lenni ahhoz, hogy a legjellemzőbb időbeni, illetve frekvenciabeli változások követhetőek legyenek, ugyanakkor kívánatos a lehető legnagyobb sűrűség elérése, hiszen a CNN-UM chip-ek véges rácsmérete miatt, a processzáláshoz szükséges idő arányos a feldolgozandó kép méretével. A cochleáris modellek a bejövő hangot általában 25, 40 csatornára bontják, de a legrészletesebb modell [83] sem használ többet 81 csatornánál, tehát a napjainkban használatos CNN-UM implementációk 64x64-es, illetve 128x128-as rács mérete vélhetően elegendően részletes frekvenciafelbontást biztosít.

A megfelelő időbeni felbontást ugyancsak a CNN rács méretének figyelembevételével kell megválasztani, az alkalmazási terület tipikus objektum (hangesemény) méretének figyelembevételével. Beszédhang esetén a zöngétlen felpattanó mássalhangzók 10 ms-nál rövidebb és a zöngés magánhangzók körülbelül 100 ms-os időtartama jelenti a két szélső korlátot. Ezen feltételek figyelembevételével az időbeni felbontást hozzávetőleg 0,5-1 ms/képpont-ra ajánlatos választani.

A fenti megfontolások alapján a dolgozat jelen fejezetében bemutatott kísérletek mindegyikében, a hangot 44.1 kHz-es mintavételi frekvenciával és 16 bites felbontással digitalizáltam. A spektrális felbontáshoz 4096 mintát tartalmazó ablakokat (W) használtam, melyeket a fent bemutatott eljárással 128 (C) csatornássá transzformáltam. A szomszédos ablakok közötti átfedést (O) 256 mintára választottam, így a kísérletekben $580\mu s/pixel$ időbeni felbontású képekkel dolgoztam, aminek következtében a 128x128-as képek 74,2 ms-nyi hangjelet reprezentáltak.

4.2. A hallási jelenet elemzés hullámszámítógépes programkönyvtára

Mint azt a frekvencia-idő kép kiszámításánál jeleztem, a mesterséges modellek többségéhez hasonlóan én sem kísérlem meg az összes csoportosítási szabály egyetlen keretrendszerbe való integrálását. Jelen esetben a harmonikus kapcsolatot azonosító csoportosítási szabály implementációja marad el. Ennek elsődleges oka, hogy a használt frekvencia-idő reprezentáció mintaillesztés-alapú megoldáshoz vezetne (lásd. 2.3.2. fejezetben), ami meglehetősen korlátozott mintáját adná az emberi hallás sajátosságainak.

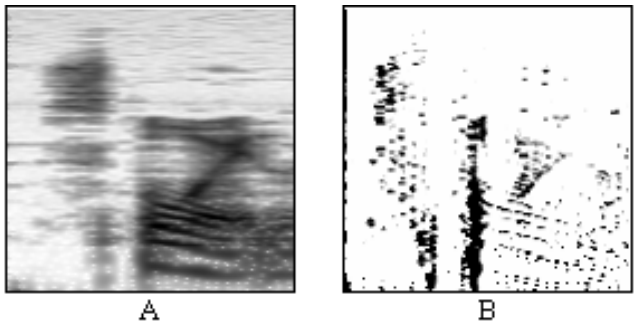
Mivel a csoportosítási szabályok, egy kivételével, a jellemző frekvencia-komponensek időbeni trajektóriáin értelmezettek, először ezek detekcióját tárgyalom.

4.2.1. A jellemző frekvencia trajektóriák detekciója

Jellemző frekvencia-komponensen a spektrogram nagy energiájú, hosszú idejű összefüggő szakaszait értem. Ezek megbízható detekciója kulcsfontosságú a későbbi csoportosítási algoritmusok sikere szempontjából. Azért, hogy a legelterjedtebb platformok [70–73] mindegyikén alkalmazható legyen, a műveletet egyszerű súlymátrix operációk szekvenciájaként adom meg. A jellemző frekvenciák kinyerése gradiens detekción alapul, aminek számítására elterjedt, robusztus megoldások állnak rendelkezésre a CNN gyakorlatban [75]. Jelen dolgozatban az alábbi súlymátrix struktúrát használom a nyugati lejtők, azaz a balról-jobbra növekvő intenzitású felületek detekciójához:

$$A = [1] \qquad B = \begin{bmatrix} 0 & 0 & 0 \\ -b & 0 & b \\ 0 & 0 & 0 \end{bmatrix} \qquad z = -bias \tag{4.5}$$

A b paraméter pontos értékét az alkalmazott CNN-UM implementációtól függően kell megválasztani, míg a *bias* paraméter a gradiens detekció érézkenységét befolyásolja. A kísérletekben bemutatott minták (4.3. ábra) $b = 2$ és $bias = 0.3$ értékekkel készültek. A fenti súlymátrix elforgatásával déli, északi, illetve a keleti lejtők azonosítása is megvalósítható.



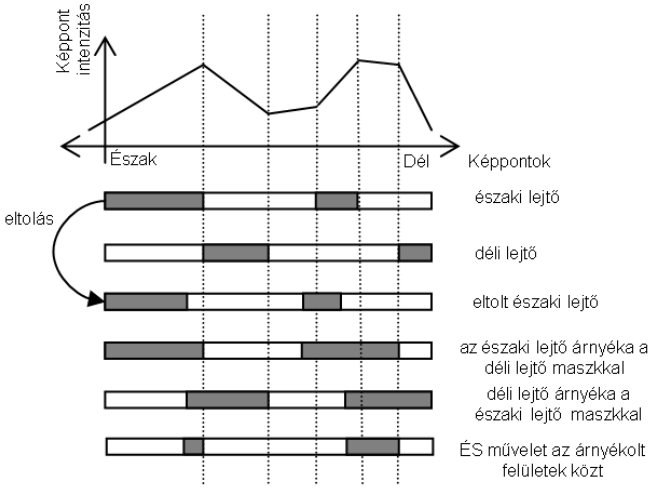
4.3. ábra. Gradiens detekció. A: bemeneti kép; B: a 4.5. egyenletként látható súlymátrixal végzett gradiens számítás eredménye.

Az ellentétes irányú lejtők között - mint az a 4.4. ábrán szemléltetett - fennsíkok vagy csúcsok találhatóak. Lévén, hogy esetünkben a keresett frekvencia trajektóriák jellemzően horizontálisak - a vertikális vonalakat széles sávú, nagy energiájú hangimpulzus okoz - a detekció a déli és az északi lejtők közötti fennsíkok és csúcsok azonosításaként értelmezhető. Mivel a déli, illetve az északi lejtőt detektáló súlymátrixokkal a dél-nyugati, dél-keleti, illetve az észak-nyugati és észak-keleti lejtők is kiemelhetőek, a nem tisztán horizontális irányú lejtők is azonosíthatóak. A **fennsík-és-csúcs** detektáló algoritmus (UML diagramja a 4.5. ábrán látható) a déli, illetve az északi gradiens súlymátrix felhasználásával a 4.4. ábrán illusztrált módon jelöli meg a fennsíkokat és csúcsokat.

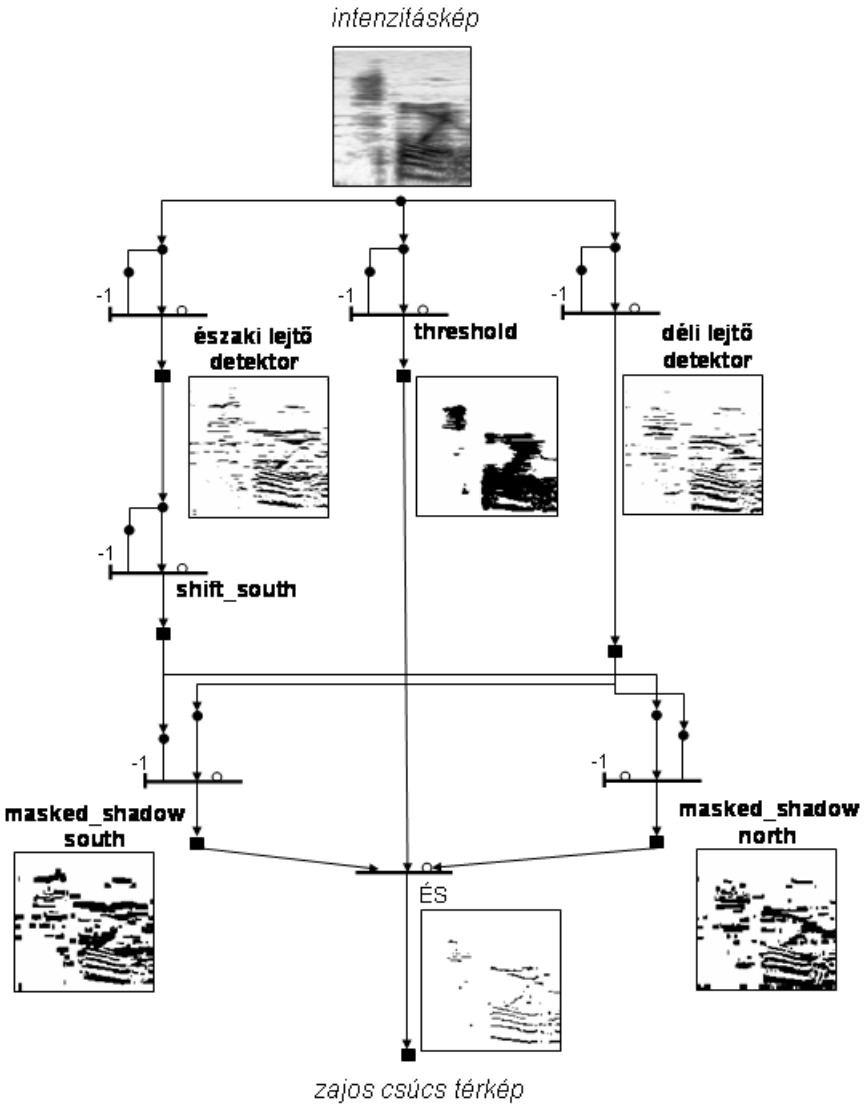
Az algoritmus első lépésében a spektrogram északi és déli lejtőinek bináris kijelölése történik meg, melyek - *északi lejtő*, illetve *déli lejtő* névvel azonosítva - közbenső eredményekként kerülnek tárolásra. A fennsíkok megjelölése a **maszkolt árnyék**[1] súlymátrix [75] alkalmazásával lehetséges, ahol az árnyékképzést - a bináris hullám terjedését - az ellentétes irányú lejtő képpontjai korlátozzák (maszkolják). Abban az esetben, ha valamelyik lejtő meredeksége túl kicsi, azaz a gradiens detekció nem jelöli meg, az árnyék terjedése a súlymátrix futási idejének korlátozásával megálllítható. Ezzel a paraméterrel egyidejűleg a legszélesebb detektált fennsík szélessége is meghatározható. Mivel a fenti operációval az egyetlen képpont szélességű csúcsok helyén nem képződik árnyék - a lejtők egymással érintkeznek, a bináris hullám a maszkkal való közvetlen érintkezés miatt nem terjed - az északi lejtőket egyetlen pixellel északra kell tolni az árnyékképzést megelőzően. A művelet végén a csúcsokat, illetve a fennsíkokat jelző pixelek képpontonként számított ÉS kapcsolattal határozhatóak meg. A folyamat egyes lépései a 4.5. ábrán követhetőek nyomon.

Mivel a digitalizálásból fakadóan a nem teljesen horizontális lejtőknek lehetnek néhány pixeles vertikális szakaszai, az ilyen módon detektált csúcsok és fennsíkok menetében apróbb szakadások fordulhatnak elő, melyek a **szakadásjavító** algoritmussal tölthetőek ki. Az algoritmus UMF diagramja a 4.6. ábrán látható. Az eljárás első lépése a végpontok, illetve a kezdőpontok kijelölése, ami a 4.7. ábrán látható **match** template-ek [75] eredményének akkumulációja révén valósíható meg.

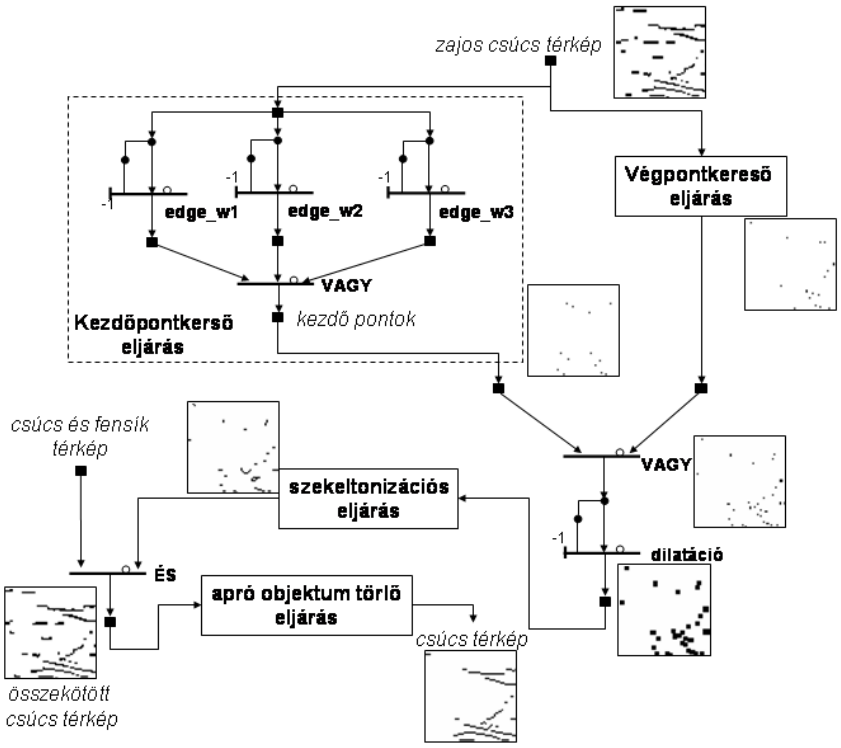
[1] masked shadow template



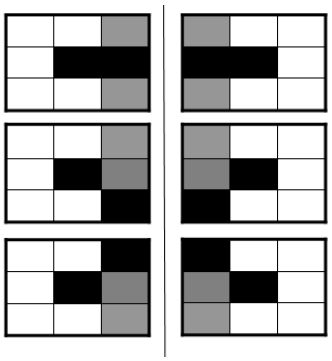
4.4. ábra. A **fennsík és csúcs detektor** eljárás működésének vázlata. Az ábra felső része egy tetszőleges térkép észak-dél irányú metszete, csúcsokkal, fennsíkokkal, völgyekkel. Az ábra alsó részén az eljárás közbenső lépései által létrehozott bináris térképek láthatók.



4.5. ábra. A **fennsík és csúcs detektor** algoritmus UMF diagramja.



4.6. ábra. A **szakadásjavító** algoritmus UMF diagramja. A **kezdőpontkereső** eljárás a szaggatott vonallal körülvett elemekként definiált. A **végpontkereső** szubrutin ennek megfelelően származtatható az alkalmazott súlymátrixok cseréjével. A súlymátrixok a 4.7. ábrán láthatóak.



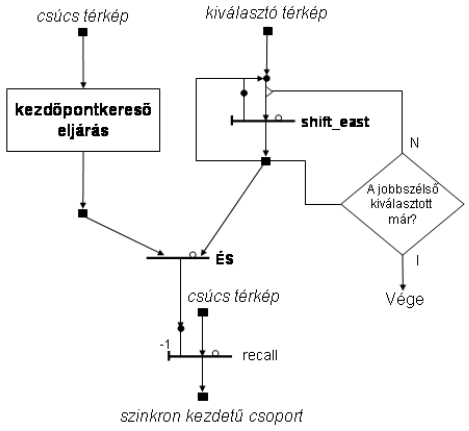
4.7. ábra. A vonalak kezdő- (bal), illetve végpontjait (jobb) megjelölő **match** súlymátrix-halmazok. A fekete, szürke, illetve fehér elemek rendre "fekete", "mindegy", illetve "fehér" elvárt értéket jelölnek.

Az egymástól egy képpont távolságban levő - a fenti módszerrel megjelölt - végződések összekötése a dilatációs template-tel [75] indított bináris hullámok összeütközésével valósítható meg. Amennyiben nem történik ütközés, a vonalak vázának meghatározásával[2] [75] a dilatáció hatása eltüntethető, míg egyéb esetekben a két végpontot összekötő szakasz marad eredményül. Ennek a kiindulási térképpel vett képpontonként számított VAGY kapcsolata a megszakításoktól mentes *csúcs térkép*, azaz a nem kívánt szakadásoktól mentes jellemző frekvencia trajektóriákat tartalmazó térkép. A későbbiekben bemutatásra kerülő algoritmusok mindegyike, ellenkező jelzés hiányában e térképet (*csúcs térkép*) használja kiindulásnként.

4.2.2. Szinkron kezdet

A természetben előforduló fizikai folyamatok által keltett hangjelenségek sajátossága, hogy spektrális komponenseik minden tagjában azonos időben jelenik meg a kisugárzott energia. Ez a tulajdonság a *csúcs térkép*en azonos időben kezdődő trajektóriákat jelent. Az alábbiakban egy olyan algoritmust mutatok be, mely bináris trigger hullámok ütközése és logikai műveletek segítségével azonosítja a különböző frekvenciasávokban megjelenő energiatartalom szinkron természetét, majd kialakítja az azonos időben kezdődő komponensekből álló hangobjektumokat.

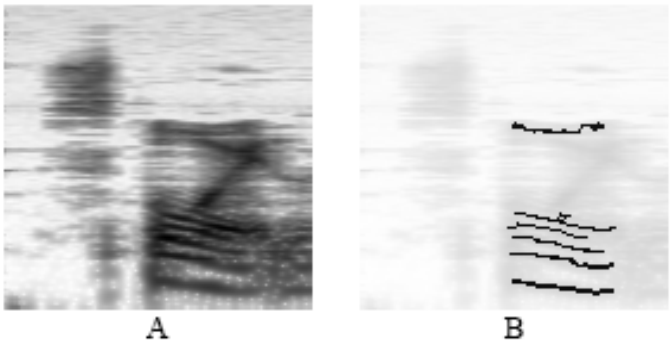
A jellemző trajektóriák kezdete a már bemutatott **kezdőpont detektor** eljárással azonosítható, a feladat tehát az így azonosított kezdő pontoknak az azonos időbeli eltolással rendelkező csoportjainak kiválasztása (a 4.8. ábrán az algoritmus UMF diagramja látható). Az azonos időben kezdődő komponenseket egy, a képet teljes magasságában kitöltő balról-jobbra haladó néhány képpont széles bináris hullámmal detektálom. A *kiválasztó térkép*en terjedő hullám minden fázisában képpontonkénti ÉS kapcsolatot hozok létre a kezdőpont kereső eljárás eredményével, ami a bináris hullám aktuális időbeni pozíciójától függő idejű szinkron kezdőpontokat eredményezi. A balról-jobbra haladó hullám szélességével a szinkron kezdőpontok közötti időbeni tolerancia szabályozható.



4.8. ábra. A **szinkron kezdet** csoportosítási szabály UMF diagramja. A **kezdőpontkereső** eljárás definíciója a 4.6. ábrán található.

[2] skeletonization

A szinkron kezdőpontok alapján végzett **recall** eljárással [75] az eredeti komponensek „visszahívhatóak" a *csúcs térkép*ről, ami az adott - a vertikális hullám aktuális helyétől függő - időpillanathoz tartozó szinkron kezdetű csoportot eredményez. Egyes CNN-UM implementációk [71, 72] képesek a sötét pontok meglétének detekciójára, így a rács tartalmának kiolvasása nélkül eldönthető van-e az adott időpillanathoz tartozó szinkron kezdetű csoport. A 4.3. fejezetben közölt futási időt a rács tartalmának kiolvasását feltételezve határoztam meg, ami erős felső becslést eredményez, mivel a CNN rácsról való adatmozgatás a VLSI környezet elhagyása miatt az egyik legidőigényesebb művelet. A tárgyalt szabály egy időpillanathoz tartozó eredménye a 4.9. ábrán látható. A balról-jobbra haladó bináris hullám 7 pixel szélességű volt, ami 4ms-os aszinkronitást engedélyez egy-egy csoporton belül.



4.9. ábra. A **szinkron kezdet** csoportosítási szabály eredménye. A.) bemeneti kép; B.) az A. kép alapján számított eredmény. A B. képen látható sziluettek demonstrációs céllal vannak feltüntetve, a szabály bináris képet eredményez.

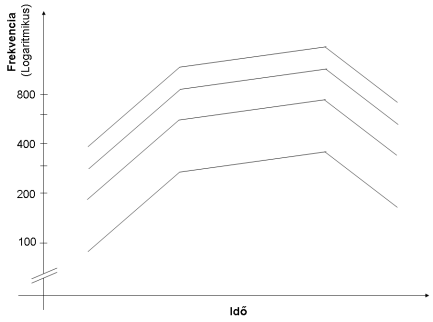
4.2.3. Közös Frekvencia-moduláció

Jelen dolgozatban közös frekvencia-moduláción (FM) több frekvencia aránytartó változását értem. Ezek ábrázolása a logaritmkus frekvencia skálán az egyes komponensek távolságtartó - párhuzamos - futásaként jelenik meg, ahol az azonos arányhoz tartozó távolságok a frekvencia emelkedésével egyre kisebbek (4.10. ábra).

Ezt a tulajdonságot használom ki a celluláris hullámszámítógépen történő detekció során a 4.11. ábrán látható blokkdiagram szerint. Először a *csúcs térkép*en levő vonalak kezdőpontjainak azonosítása történik meg a már bemutatott módszerrel.

Annak érdekében, hogy ezeket a vonalakat egymástól elkülönítve kezelhessem, a kezdő pontokat digitális processzoron tartom nyilván. A közölt modellben ez az egyetlen lépés, ahol a digitális számítási paradigma alkalmasabb volta miatt érdemes elhagyni a CNN-UM chip felületét. Természetesen a chip-ekre integrált belső memória korlátozott mennyisége miatt, az alkalmazott implementációtól függően egyéb esetekben is szükség lehet a rácson levő képek elmentésére, ez azonban nem az architektúra és a feladat kapcsolatából fakadó következmény.

A digitális számítógépen (Bi-I implemetációk esetén a platform-on elhelyezett DSP-n) tárolt pontok közül kiválasztunk egy pixelt, amelyhez tartozó görbét a **recall** operációval [75] a CNN



4.10. ábra. Az közös frekvencia-modulációjú komponensek párhuzamosokként jelennek meg a logaritmikus frekvencia skálán ([6] nyomán).

4.11. ábra. A közös FM csoportosítási szabály UMF diagramja. A **párhuzamos görbe kereső** szubrutin a 4.12. ábrán látható.

rácsra rajzolunk, majd megkeressük az ezzel állandó távolságban levő egyeneseket, melyek közös FM csoportot alkotnak.

Minden közös FM csoport létrehozását követően pixelenként kizáró-VAGY kapcsolattal töröljük a csoportba tartozó görbéket a *csúcs térkép*ről, majd a fennmaradó görbék kezdőpontjait felhasználva frissítem a közös FM csoportba még nem sorolt görbék kezdőpontjait tartalmazó listát. Amennyiben egy, vagy kevesebb pont van a listában, az összes közös FM csoportot azonosítottuk.

Párhuzamos görbének tekintem azokat, melyek pontjainak túlnyomó részének egy kiválasztott (referencia) görbétől mért vertikális távolsága adott $N_1 - N_2$ határon belül marad. Mint az a 4.12. ábrán látható, először a kiválasztott görbéhez $N_1 - N_2$ távolságra levő sötét pixelek kiválasztása történik meg, melynek eredménye adja az un. *párhuzamos képpont* térképet.

A *párhuzamos képpont* térképről csak azokat a képpontokat hagyjuk meg, melyek egy majdnem teljes görbét alkotnak, azaz képontjainak elenyésző része esik ki a képpont keresés tartományából. Ennek az elenyésző résznek a mérete az **apró objektumok eltüntetése**[3] [75] szubrutin iterációinak számával kontrollálható. Az adott sötét képpontól N_m ($N_1 <= N_m <= N_2$) távolságra levő sötét pontok megtartását végző NxN-es template osztály létrehozása a 3.2. állapotegyenletet felhasználva megfelelő robusztussággal megalkotható, mely általános alakban az alábbi formában írható:

$$A = 2 \qquad B = \begin{bmatrix} 0 \\ n_m \\ 0 \\ n_m \\ 1 \end{bmatrix} \qquad z = -1 \tag{4.6}$$

ahol n_m egy $N_m - 1$ dimenziós null vektort jelöl. A fenti template osztály ($N_1 <= N_m <= N_2$) távolságokra adott eredményének pontonkénti VAGY kapcsolatával megkaphatóak az ($N_1 - N_2$) tartományban levő sötét pixelek.

A bemutatott NxN-es template szilícium alapú CNN-UM [71, 72] implementációkon nem futtatható, azonban mivel bináris morfológiai operációt valósít meg, ismert dekompozíciós módszerek állnak rendelkezésre [82, 84]. A template ritkásságának köszönhetően - kevés nullától különböző elemet tartalmaz - $N_2 - 1$ lépésben dekomponálható a 4.13. ábrán látható UMF diagram segítségével. Az említett template-osztály 3x3-as verziójának robusztussági értéke meghaladja a 0.7-es értéket, tehát szilícium alapú CNN-UM implementációkon is sikerrel alkalmazható. A dekompozíció során használt egyéb template-ek szilícium felületen való sikeres futtatását gyakorlati alkalmazások igazolják [82].

A 4.14. ábra az előbb részletezett lépeseket illusztrálja. Az ábra A részén látható a kiválasztott görbe, míg a B kép a lehetséges görbék halmazát, azaz *csúcs térképet* szemlélteti. Az ábra C része a párhuzamos képpontok térképét ábrázolja.

Az $N_1 - N_2$ távolságtartományban levő párhuzamos görbék azonosításán túl meg kell találnunk az egyéb távolságtartományban levő görbéket is, ezért az eljárást ismételten végre kell hajtani az $N_1 - N_2$-től különböző tartományokra (4.12. ábra). Ebben az esetben azonban már az előzőleg megtaláltakkal párhuzamosak görbéket is megtaláljuk, azaz *kiválasztott görbe* térképen tartjuk az előző iterációk eredményét. A végső eredmény, az egyes iterációk VAGY kapcsolataként áll elő, a 4.15. ábrán szemléletett módon.

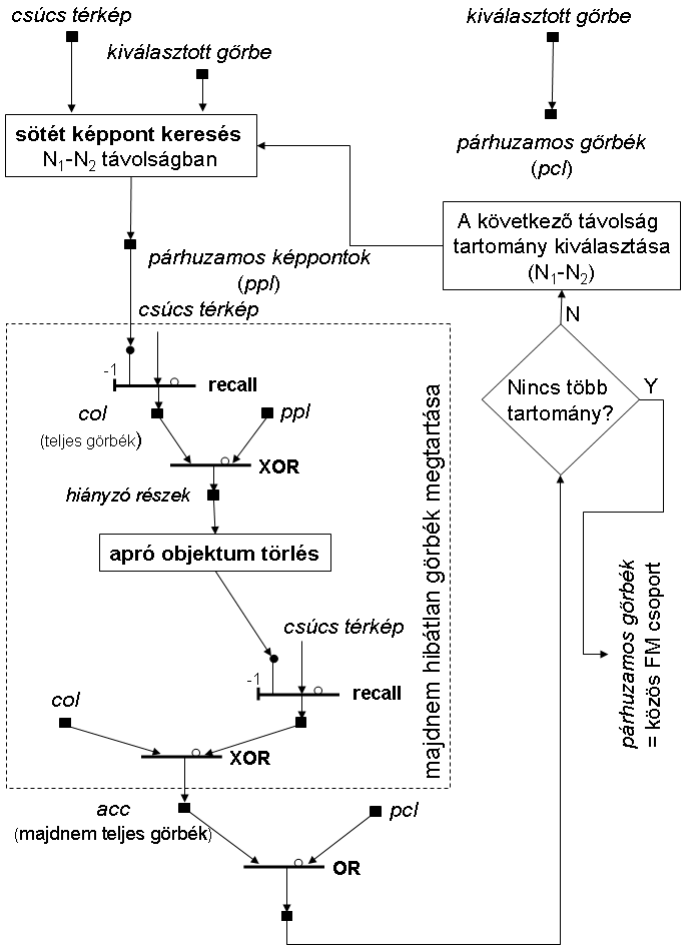
Annak eldöntésében, hogy mely és mekkora távolságtartományokban szükséges a keresést elvégezni, a 2.1.1. fejezetben említett pszichoakusztikai tapasztalatok lehetnek segítségünkre. Első-

[3] small object removal

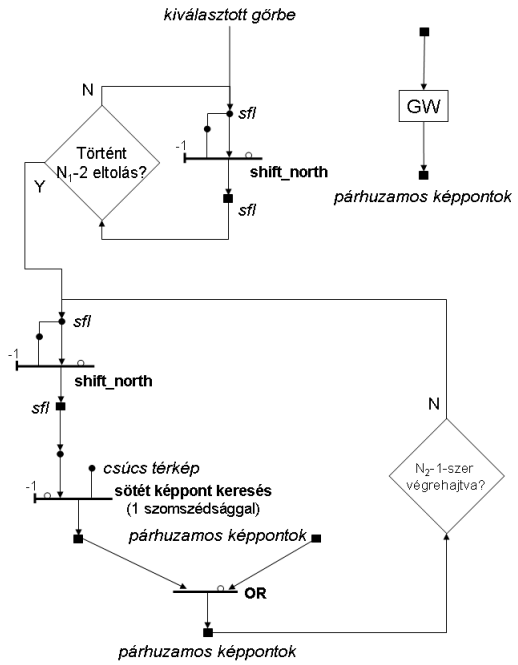
Alkalmazott keresési távolságok [képpont]
11-16
10-16
11-16
10-15
8-12
6-10
8-12
6-10
6-10
4-7
2-4
10-15
8-12
6-10
8-12
6-10
6-10
4-7
2-4

4.1. táblázat. A közös FM csoport létrehozásakor használt keresési távolságok.

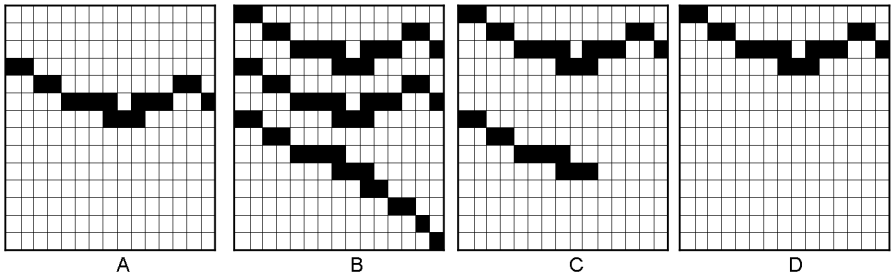
sorban az egymáshoz közeli frekvenciákon levő komponensek csoportosítása a valószínű, ennélfogva felső határ adható a keresési távolságra, sőt mivel a komponensek a frekvencia emelkedésével egyre közelebb kerülnek egymáshoz az egyes görbék, a keresési távolság folyamatosan csökkenthető. A keresési távolságtartományok szélességének megválasztása egy optimalizációs folyamat eredményeként határozható meg, melyben a hibásan detektált párhuzamosok, illetve a nem detektált ám közös FM csoportba tartozó görbék számának minimalizálása a feladat. A közölt kísérletekben a maximális keresési távolságot 15 képpont nagyságúra választottam és a 4.1. táblázatban látható heurisztikusan meghatározott távolságtartományokon végeztem el a párhuzamos görbe keresést.



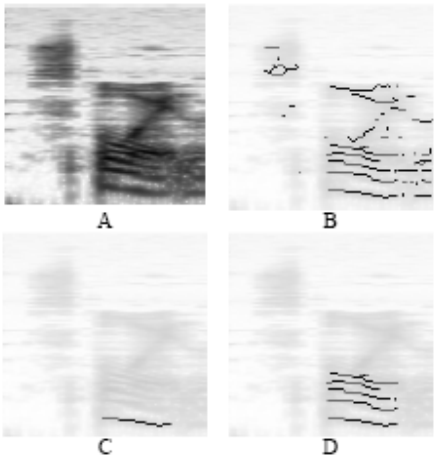
4.12. ábra. A közös FM csoportok létrehozásához használt **párhuzamos görbe keresés** UMF diagramja. Az eredmény a *kiválasztott görbé*vel párhuzamos görbék térképe. A létrehozott közös FM csoport az itt kiválasztot görbéket tartalmazza. Az $N_1 - N_2$ távolságban levő sötét képpontok az N_m ($N_m = N_1 \ldots N_2$)távolságban megtalált képpontok VAGY kapcsolatával előállítható, illetve a szilícium implementációkkal kompatibilis csak 3x3-as template-eket alkalmazó verzió a 4.13. ábrán látható.



4.13. ábra. Az $N_1 - N_2$ távolságban levő sötét képpontokat kereső algoritmus 3x3-as template szekvenciára való felbontásának UMF diagramja. [82] alapján, a dekomponálandó template-osztály egyetlen nem nulla eleméből adódó egyszerűsítések figyelembevételével.



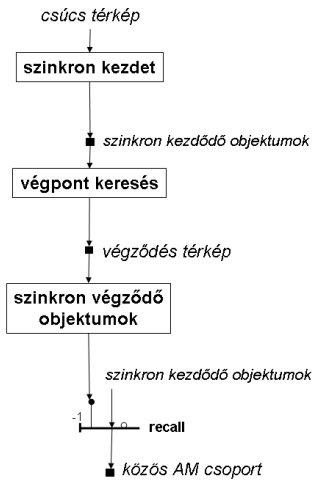
4.14. ábra. A **párhuzamos görbe keresés** szubrutin egyes lépéseinek illusztrációja. A.) a *kiválasztott görbe* térkép. B.) a rendelkezésre álló görbe halmaz (praktikusan a *csúcs térkép*), melyekből kiválasztjuk az A. képen láthatóval pérhuzamosakat. C.) A 3 képpont távolságra elvégzett **párhuzamos képpont keresés** eredménye. D.) A 4.12. ábrán definiált **majdnem hibátlan görbék megtartása** eljárás eredménye. Ezen térkép A képpel számított képpontokénti VAGY kapcsolata adja a közös FM csoport egy komponensét.



4.15. ábra. A közös FM csporotosítási szabály eredménye. A.) A bemeneti spektrogram; B.) A *csúcs térkép*; C.) Egy *kiválasztott görbe* térkép; D.) A C-vel párhuzamosnak talált görbék halmaza, ami egy közös FM csoportot alkot. Természetesen a *csúcs térkép* egyéb alkalmas mérettartományba eső görbéire az itt látotthoz hasonlóan el kell végezni a közös FM csoport keresést, de ez ebben az esetben nem eredményezett újabb csoportot.

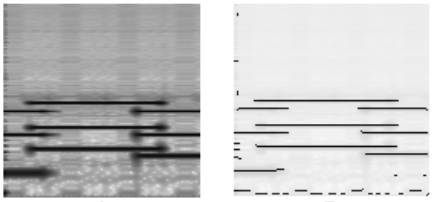
4.2.4. Közös Amplitúdó-moduláció

Közös AM csoportoba tartoznak azok a komponensek melyek amplitudó változása - megszűnik, illetve megjelenik - időben szinkron módon történik. Az azonos időben megjelenő, illetve eltűnő komponensek megtalálásának kérdése részben visszavezethető a már megoldott problémákra (lásd 4.16. ábra).



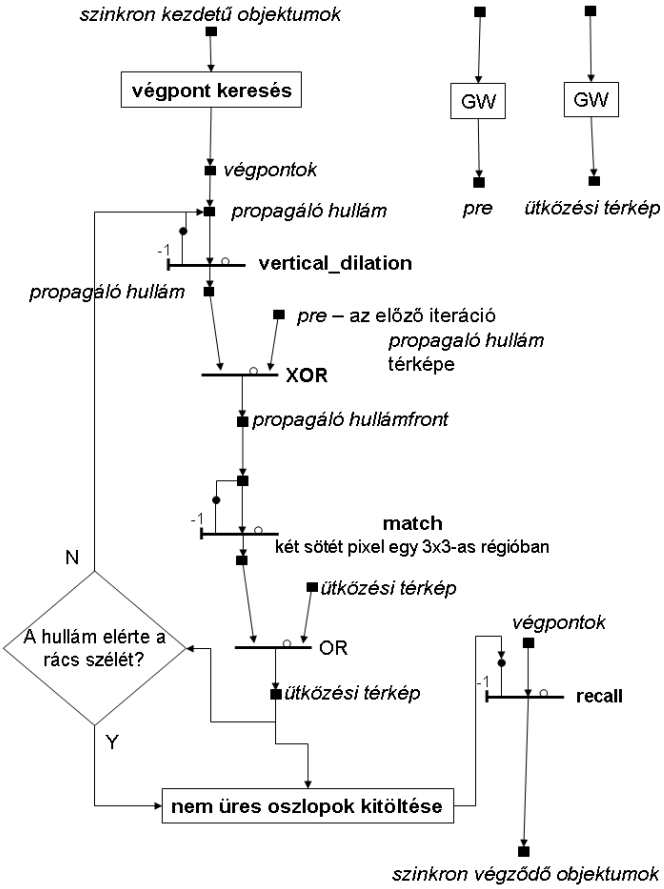
4.16. ábra. A közös AM szabály UMF diagramja.

A 4.2.3. fejezetben bemutatott módszerrel először kiválasztom a szinkron kezdődő komponenseket, majd ezek közül kijelölöm azokat, amelyek azonos időben érnek véget. Az azonos időben való végződés azonosítását ismét bináris hullámokkal való ütközés révén valósítjuk meg. A 4.19. ábrán látható módón az algoritmusnak a **szinkron kezdet** csoportosítási szabály eredménye a bemenete, amin az objektumok végpontjainak azonosítását követően néhány pixel széles, vertikálisan lefelé és felfelé is terjedő hullámokat indítok a végpontokból. Amennyiben ezek a hullámok ütköznek - a detektált végpontok időbeni eltérése a vertikálisan terjedő hullám szélességénél kisebb - tudható, hogy a szinkron kezdődő objektumok azonos időben is végződtek (lásd 4.19. ábra). A végpontok és a vertikális hullám ütközésének pontjai alapján végrehajtott **recall** operációval közös AM csoport hozható létre. A 4.17. és 4.18. ábrák az algoritmus eredményét mutatják. Az ábrák elkészítéséhez a szinkron végpontok és kezdőpontok időbeni eltérését 7 pixel széles vertikális hullámok/egyenesek terjedésével azonosítottam, ami az egyes komponensek közt maximum 4ms-os eltérést engedélyez.



4.17. ábra. A közös AM csoport szemléltetésére létrehozott hangból készített spektrogram. Az *A* képen 3 harmonikus komplex spektrogramja látható. Az A kép bal oldali hangja a 200Hz-es alapfrekvencia 8. és 12. felharmonikusaiból áll; a középső: 500Hz-es alapfrekvenciájú a 2. és a 4. felharmonikust tartalmazó hang; a jobb szélső: 400Hz alapharmonikus 2-szeres és 4-szeres komponenseként jött létre. Az ábra B részén az A képből számított *csúcs térkép* látható.

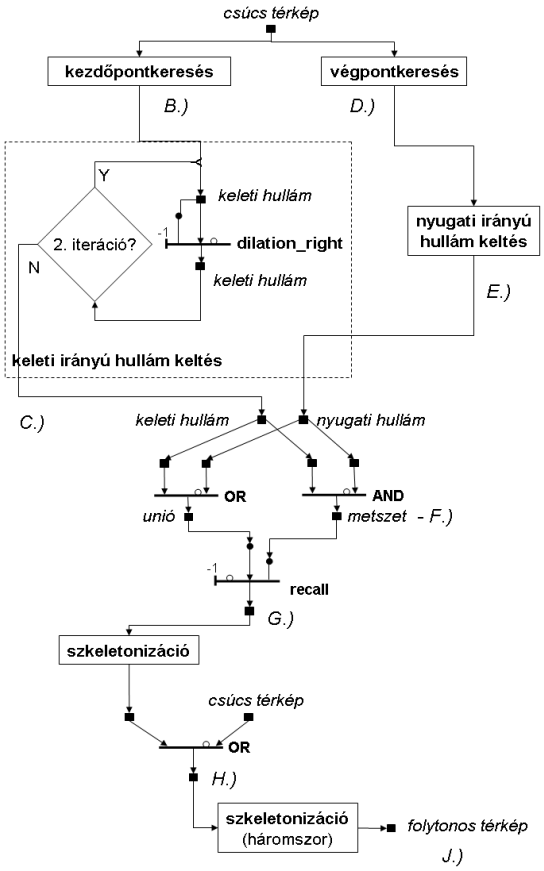
4.18. ábra. A 4.17. ábra alapján létrehozott 3 közös AM csoport.



4.19. ábra. A szinkron végződő objektumokat azonosító algoritmus. A **nem üres oszlopok kitöltése** szubrutin egy a rács szélességétől függő alkalommal végrehajtott vertikális hullám terjesztés, mely kitölti a szinkron végződéseket tartalmazó oszlopot.

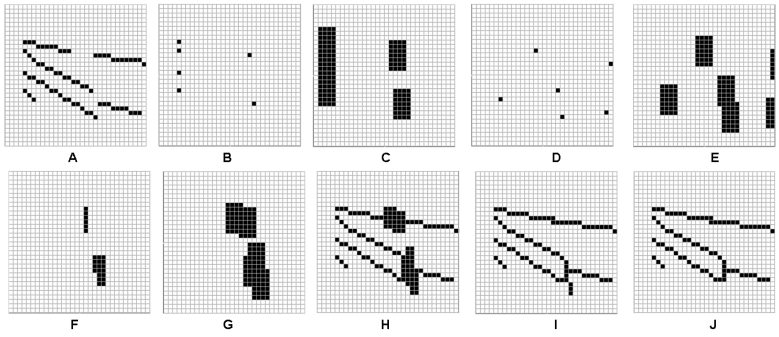
4.2.5. Folytonosság

Előfordul, hogy a hangforrások a kisugárzott hangenergiát rövid megszakítást követően egy, az addigi frekvenciához közeli sávban sugározzák tovább, melyek az alkalmazott frekvencia-idő reprezentációban rövid „résekként" jelentkeznek. A hullámszámítógépes megvalósítás ennek megfelelően célirányos és némi hasonlóságot mutat a **szakadásjavító** algoritmussal (4.6. ábra), jóllehet ebben az esetben kifejezetten végpontok kezdőpontokkal való összeköttetése a cél, míg a **szakadásjavító** eljárás esetén csupán a rácson bizonyos távolságon belül levő pontok összekötése a feladat.



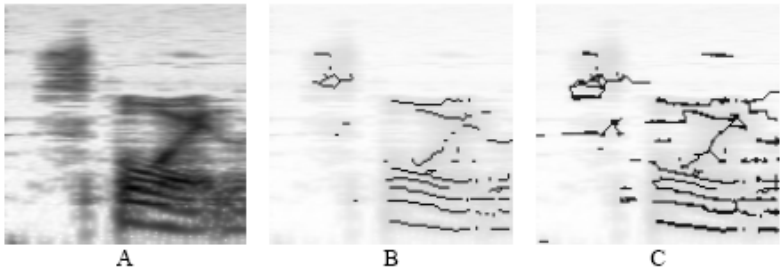
4.20. ábra. A folytonosság csoportosítási szabály UMF diagramja. A nyugati, illetve keleti irányú hullámokat keltő szubrutinok azonosak, eltekintve attól, hogy a keleti irányú bináris hullámok képzésénél a **dilation_right** [75] template használatos. Az egyes operációk eredményénél feltüntetett egy betűs kód (*A.)* - *H.)*), a 4.21. ábra megfelelő részeit azonosítja.

A 4.20. ábrán látható módon először a vég- és a kezdőpontok detekciója történik meg, amelyekből nyugati, illetve keleti irányba terjedő hullámokat indítunk. Abban az esetben, ha az ellenkező irányba terjedő hullámok ütköznek - a két térkép ÉS kapcsolata tartalmaz sötét pixelt - az érintett



4.21. ábra. A folytonosság csoportosítási szabály hullámszámítógépen való implementációjának lépéseit szemléltető ábra. A.) A kezdeti kép (a *csúcs térkép*) melyen a szabályt alkalmazni kívánjuk; B.) kezdőpontok; C.) a kezdőpontokból induló nyugatra terjedő hullámok; D.) végpontok; E.) a végpontokból keletre propagáló hullámok; F.) az E és C hullámok metszete; G.) az ütköző (nem üres metszetű) hullámok, a regenerálást követően; H.) az A és a G képek VAGY kapcsolatánból képzett eredmény egy szkeletonizációs lépést követően; I.) a szkeletonizáció végső eredménye; J.) a szkeletonizáció által létrehozott nem kívánt végződések pruning-el [75] történő eltávolítást követően.

kezdő és végpont között a megadottnál kisebb távolság van, tehát összeköttetést kell létesíteni. Az összeköttetés megvalósítása a **szakadásjavító** eljáráshoz hasonló módon a szkeletonizáció [75] révén valósítható meg a 4.21. ábrán látható módon. Az algoritmus a fejezet elején bemutatott példán futtatva a 4.22. ábrán látható eredményt adja. A példán 3-3 iterációban terjesztettem a jobbra, illetve balra terjedő hullámokat, ezzel maximum 5 pixel horizontális szélességű rések betömését valósítottam meg, ami az egymástól 2.5ms-nál rövidebb ideig tartó szakadással bíró hangobjektumok csoportotsítását teszi lehetővé.

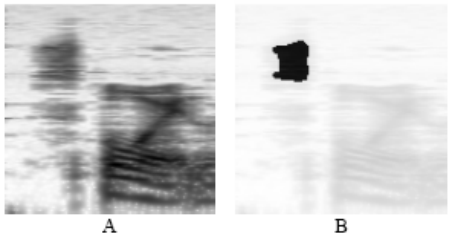


4.22. ábra. Példa a folytonosság szabályra. A.) a spektrogram; B.) a *csúcs térkép*; C.) A folytonosság csoportosítási algoritmus eredménye maximum 4 képpontnyi távolság kitöltése esetén.

4.2.6. Közelség

E szabály esetén az előzőekkel ellentétben nem a *csúcs térkép*et használom bemenetként, mivel azon a szabály által csoportosítani kívánt sűrűn elhelyezkedő, nagy energiájú, tüske-szerű komponensek

nem képeznek trajektóriákat. Ennek ellenére a szabály az eddigieknél is jobban alkalmazkodik a celluláris számítási környezethez, mivel kis területen nagy energiájú komponensek detekciója a cél, ami egyetlen lépésben megvalósítható az **average and threshold**[4] template-tel [75]. A template futásának eredményeként kijelölhetőek azok a területek, ahol a megadott *bias* áramnak megfelelő küszöböt túllépi a template területén belüli energia átlag. Az így kijelölt területeket nagyságuk szerint szűrni kell, melynek paramétereit az alkalmazástól függően kell beállítani. A 4.23. ábrán látható eredményt a **small object removal** [75] eljárás háromszori futtatását követő **recall** operáció eredményeként kaptam. Az algoritmus egyszerűsége miatt ezúttal eltekintek az UMF diagram közlésétől.



4.23. ábra. Példa a közelség csoportosítási szabály által egy objektummá alakított 'sz' hangra.

A *bias* áram értéket a zajnak, a jelszintnek, illetve az átlagos energia tartalomnak a figyelembevételével lehet meghatározni. Az átlagos jelenergia meghatározására a legmodernebb CNN chip-ekben speciális ellenállás-hálózatok állnak rendelkezésre, melyek segítségével az energia átlag egy dilatációs tranziens segítségével megkapható.

4.3. Futásidő analízis

A hullámszámítási paradigma alkalmazásának elsődleges indoka a feladat és a CNN architektúra szerencsés összhangjából adódó alacsony számítás igény, rövid futási idő. Ezt alátámasztandó megvizsgáltam a bemutatott csoportosítási szabályok várható teljesítményét. Az analóg megvalósítások közös jellemzője, hogy a tulajdonképpeni számítás - az tranziensek és a logikai műveletek - végrehajtásához szükséges idő néhány 10 ns-os nagyságrendben mozog. Ennél egy nagyságrenddel időigényesebb a celluláris tömbről való adatkiolvasás, ezért mint azt a 3.4. fejezetben már említettem az algoritmusok tervezésénél törekedtem ennek elkerülésére. Jelen programkönyvtár futtatásakor egyetlen esetben kényszerültem a tömbről való adatkiolvasásra, azonban a napjainkban hozzáférhető szilícium alapú CNN-UM megvalósítások korlátozott lokális memória-kapacitása miatt (a [72] esetén négy analóg (LAM) és négy bináris (LLM) lokális memória érhető el) természetesen egyéb esetekben is szükség lehet az adatok kiolvasására, ami drasztikus hatással van az összetett műveletek számítási idejére. A fentiek miatt a 4.2. táblázatban közölt futási idő adatokat a [72]-ben közölt analóg chip paraméterei alapján számítottam ki, feltételezve azt, hogy csupán két analóg és két logikai memória áll rendelkezésre. Feltételeztem, hogy az algoritmusok futtatásához szükséges template-ek az algoritmusok inicializációs fázisában a CNN chip-re tölthetőek, így azok futtatása előtt nincs szükség ezek feltöltésére. A könyvtár néhány elemének futás idejét az InstantVision EyeRis [85] 2.3.4-es verziójának segítségével megmértem a Bi-I v2. platformon. A mért adatok több

[4] avertrsh

Művelet	Becsült futási idő	Mért futási idő
Jellemző frekvenciák kinyerése	1.91 ms	3.57 ms
Közös kezdet/vég	5.23 ms +0.135 ms / közös kezdet/vég csoport	-
Közös FM	0.775 ms + 5.62ms / közös FM csoport	-
Közös AM	5.23 ms + 6.05 ms / közös AM csoport	-
Folytonosság	3.68 ms	-
Közelség	1.72 ms	2.85 ms
Összeg	**20.4 ms (54.4 ms)**	-
Minta alkalmazás	**74 ms**	-

4.2. táblázat. A Bi-I v2 [72] és a Bi-I v2 paraméterevel megegyező, ám végtelen template memóriával, két LAM-al és két LLM-el ellátott teoretikus CNN-UM implementáció mért, illetve becsült számítási teljesítménye. A zárójelben közölt összeget minden csoportosítási szabályból három-három objektumot feltételezve számítottam. Ez a spektrogramon ábrázolt hangjel hosszúsága (74 ms - a minta alkalmazásra vonatkozó idő) miatt a gyakorlatban előforduló maximális hangobjektumsűsűségnek tekinthető, ami igazolja a valós idejű alkalmazás lehetőségét. A mért és a becsült értékek közti eltérés a használt template-ek chip-re töltéséből származó többlet időből, valamint az eltérő mennyiségű *onchip* memória kapacitásból adódik.

iteráció átlagolt eredményei melyek nem tartalmazzák a hálózati forgalomból származó adatátviteli többletet. A köztes és a végleges eredményeket csak a Bi-I platform memóriájában tároltam el.

Mind a becsült, mind a mért eredmények igazolják, hogy a hullámszámítási paradigma segítségével az algoritmusok valós időben futtathatóak, azonban, mint azt az eredmények mutatják, a jövőben fontos optimalizációs szempont lehet a CNN implementációk chip-re integrált memória mennyisége. Az algoritmusok komplexitása elérte azt a szintet, amikor a template-ek és a köztes eredmények tárolása, illetve mozgatása miatt fellépő I/O többlet jelentősen befolyásolja az elérhető teljesítményt.

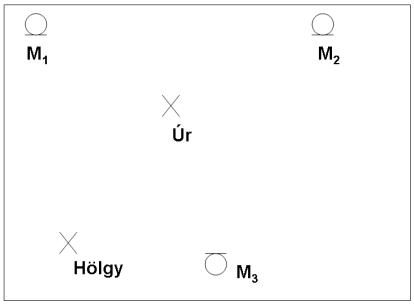
4.4. Alkalmazási példa

Mint azt a fejezet bevezetőjében említettem jelen dolgozatban nem vállalkozom a primitív csoportosítási szabályok kimenetének magasabb rendű folyamatok által vezérelt újraértelmezésére, azonban az alábbiakban egy alkalmazási példát mutatok a "hullámszámítógépes hallási jelenet elemzés programkönyvtár" felhasználására.

Mint az az 5. fejezetben olvasható a forrás-lokalizáló algoritmusok teljesítményét nagyban befolyásolja a jelek periodicitása, sőt több hangforrás jelének elegye esetén az algoritmusok által szolgáltatott forráshely-becslés kiszámíthatatlan eredményre vezet, mivel a jelek kereszt-korrelációjából számított idő-különbség nem határozható meg egyértelműen. Különösen igaz ez azokban az egyébként hétköznapi esetekben, amikor nincs információnk arról, hogy a rögzített elegy hány egyidejűleg sugárzó forrás jelének összegeként áll elő.

A probléma megoldására több munkában tettek kísérletet, részben a beszélők alapfrekvenciájának, illetve az azokhoz tartozó felharmónikusok meghatározásával [27, 28], részben a zöngétlen

mássalhangzók detekciójával [23,24]. Az általam bemutatott módszer ez utóbbiak példáját követve a zöngétlen mássalhangzók kezdetét detektálva határozza meg a beszélők helyét. A kísérletben



4.24. ábra. A forrás lokalizáló kísérlet elrendezése. A mikrofonok a (5.8m ,0.1m), (1.05m ,0.1m), (3.7m ,5.1m) koordinátájú helyeken kerültek elhelyezésre.

egy férfi és egy nő hangjának elegyét rögzítettem a 4.24. ábrán látható szimulált és visszhangmentes környezetben elhelyezett mikrofonokkal. A beszélők nem mozognak, azonban amennyiben mozgásuk sebessége jelentősen elmarad a jelfeldolgozás, illetve a hangterjedés sebességénél nem okoz problémát. (Egyéb esetekben rövidebb jelszakaszból kell a kereszt-korrelációt számítani, illetve kompenzálni kell a Doppler effektus hatására fellépő frekvencia-tartalomváltozást.) Az egyes beszélők hangjából, illetve azok elegyéből képzett spektrogramok a 4.25. ábra A, C, illetve E részein láthatók. A rögzített jelek 8192 minta hosszúságú szegmenseit felhasználva forrás-lokalizációt végeztem a [86]-ban közölt, és jelen dolgozat 5.4.4. fejezetében ismertetett. Első esetben a lokalizációra használt jelszegmenseket valamely véletlen időpillanattól kezdve állandó időközönként választottam ki, ügyelve arra, hogy legalább az egyik beszélő aktív legyen. A másik esetben az 1. mikrofon által rögzített jel cochleáris transzformáltján az előzőekben bemutatott **közelség** csoportosítási szabály felhasználásával kiválasztottam a leginkább a zöngétlen beszéd-szegmenseknek megfelelő hangobjektumokat (lásd 4.25. ábra B, C, D részei.), majd a **szinkron kezdet** csoportosítási szabályt alkalmazva hang objektumokká formáltam a zöngétlen hangokat jelző "foltokat". Az így azonosított 15 hang objektum látható a 4.26. ábrán. Az azonosított hang objektumok nyugati gradiensének detekciójával meghatároztam azok kezdőfrontját, majd a kezdőfront időbeni elhelyezkedése és spektrális kiterjedése alapján mindhárom mikrofon jeléből kiválasztott 8192 minta hosszúságú szegmenst felhasználva végeztem el a forrás helyének meghatározását.

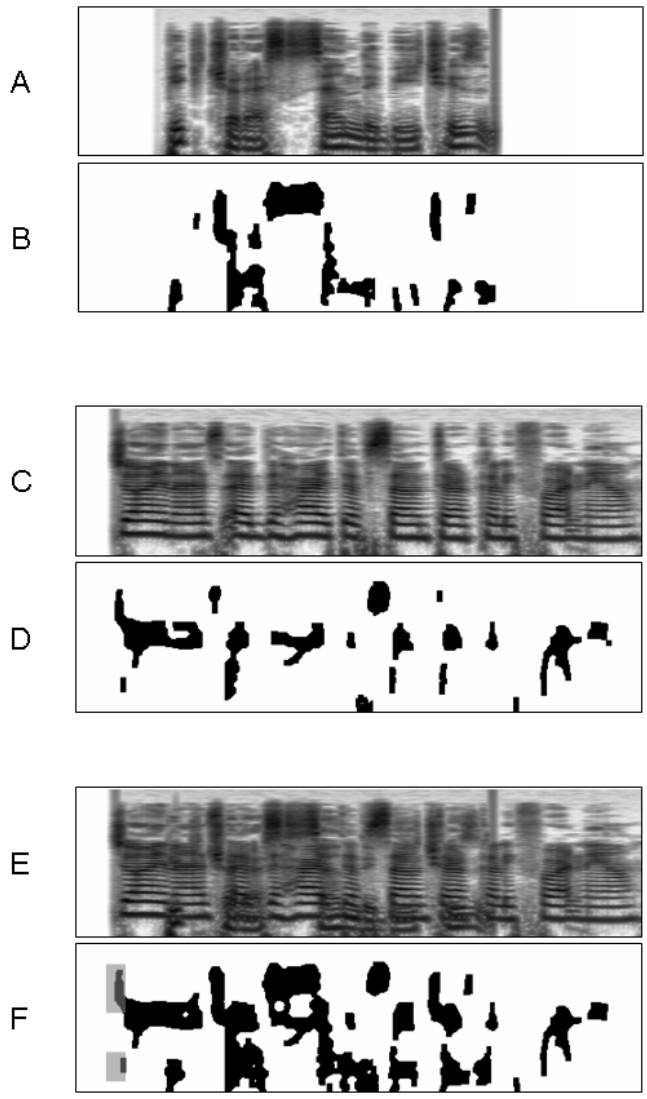
A forrás-lokalizáció teljesítményét az első esetben - mivel nem tudható, hogy a felhasznált szegmens éppen melyik beszélőhöz tartozott - az algoritmus eredményének legközelebbi beszélőhöz viszonyított távolságaként értelmeztem, míg a második esetben, amennyiben eldönthető, hogy melyik beszélőhöz tartozott a kiválasztott szegmens, a jelet kibocsátó beszélő helyéhez mérten határoztam meg a lokalizáció hibáját. Azokban az esetekben, amikor "foltok" átfedése miatt nem határozható meg a szegmens tulajdonosa, hasonlóan az első eljáráshoz, a legrövidebb távolságot vettem a forrás-lokalizáció hibájául. Az azonosított hangobjektumok beszélőkhöz rendelése a 4.26. ábra alapján manuálisan történt meg. A kapott eredmények a 4.3. táblázatban láthatóak. A fenti teljesítmény-ellenőrzési kritériummal biztosítottam, hogy a hagyományos módszer a lehető legkisebb hibával teljesítsen, szemben azzal, amikor a hangobjektum "tulajdonosához" mért távolságot vettem a lokalizáció hibájául.

Szegmens sorszám	A lokalizáció hibája a bemutatott módszerrel választott jelszegmenseken		A lokalizáció hibája véletlenszerűen választott jelszegmenseken. [m]
	beszélő	Lokalizációs hiba [m]	
1	hölgy	2.328	1.607
2	úr	0.056	0
3	nem eldönthető	0.039	0.08
4	nem eldönthető	0	0
5	hölgy	0.019	0
6	hölgy	1.004	2.474
7	nem eldönthető	1.498	0.02
8	úr	0.019	1.584
9	nem eldönthető	0	1.811
10	úr	0	0.04
11	úr	0	1.498
12	hölgy	0.019	0.02
13	hölgy	0.044	1.533
14	hölgy	0.551	2.6
15	hölgy	0.019	1.721
Átlagos lokalizációs hiba	**0.373**		**0.999**
Szórás	**0.701**		**0.994**

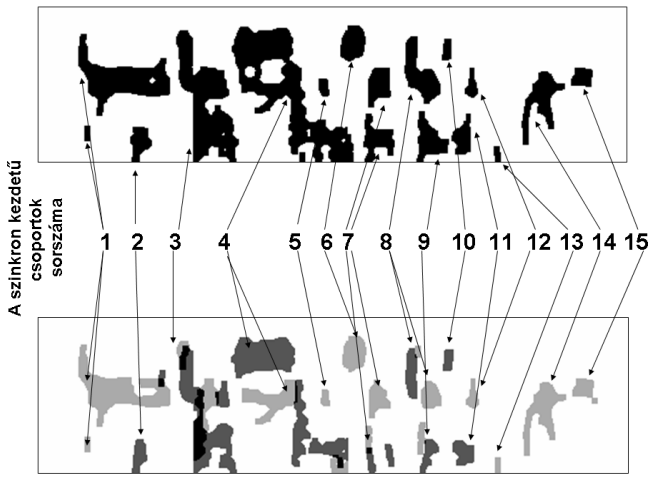
4.3. táblázat. A lokalizációs hiba a két módszerrel kiválasztott 15 jelszegmens esetén.

Mint azt az eredmények igazolják, a bemutatott *hallási jelenet elemzés* könytár segítségével kiválasztott jelszegmenseket felhasználva a lokalizációs hiba közel harmadára csökkenthető.

A bemutatott példa természetesen csak korlátozott modelljét adja a valóságban előforduló körülményeknek. Nyilvánvaló, hogy a felsőbb sémavezérelt szabályok segítsége nélkül csak néhány, egy időben jelenlevő forrás esetén vagyunk képesek az egymással átfedő jelek szegregációjára, tehát több beszélő, vagy más háttérzajforrás jelenléte esetén nem minden esetben érnénk el ilyen jó eredményt.



4.25. ábra. A férfi (A) és a női (C) hang spektrogramja, valamint az ezeken végzett *közelség* algoritmus eredménye (B,D). Az ábra E része az M_1 mikrofon által rögzített elegy spektrogramját ábrázolja, míg az F kép ez utóbbi spektrogramon végzett *közelség* eredménye. Az F ábra bal szélén szürkével jelöltem az első objektum helyének meghatározásához használt jelszakasszal ekvivalens területet.

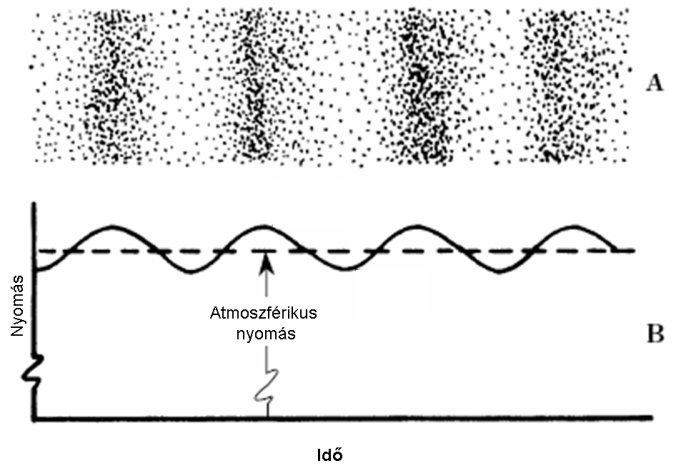


4.26. ábra. A szinkron kezdetű csoportok a közelség szabállyal képzett eredményen (felül). A nyilak a *szinkron kezdet* csoportosítási szabály alapján azonos csoportba sorolt objektumokat jelzik. Az ábra alsó részét a 4.25. ábra B és D részeinek összegeként hoztam létre, azért, hogy követhető legyen, hogy az azonosított hang objektumok, melyik beszélő hangjának az eredményei. (világos szürke: női, sötét szürke: férfi, fekete: közös) Mint az az ábrán látható a 3., 4., 7. és 9. csoportok kezdőfrontja mindkét beszélő hangját tartalmazza, ezért ezekben az esetekben a hozzárendelés nem elvégezhető.

Napjaink hangforrás-lokalizáló algoritmusai

5.1. A hang mint fizikai hullám

Fizikai jellegét tekintve a hang valamilyen rugalmas közeg mechanikai rezgése, ahol a rugalmas anyag azon részecskéi, amelyek külső hatásra kimozdulhatnak nyugalmi helyzetükből, a rugalmassági erő és a tehetetlenség folytán periódikus rezgésbe jönnek (lásd 5.1. ábra). A levegőben terjedő



5.1. ábra. A hang terjedése. A.) a levegő részecskék elhelyezkedése a hullámban; B.) a légnyomás alakulása

hang légnyomásingadozás formájában terjed. Az állandó értékűnek tekinthető légköri nyomásra szuperponálódik a hangnyomás ($p(t)$). A tér egy pontjában az eredő $P(t)$ légnyomás a P_0- al jelölt

konstans légköri nyomás és a $p(t)$ összegeként adható meg.

$$P(t) = P_0 + p(t) \tag{5.1}$$

Információt a hang időbeni változása hordoz, ennek megfelelően csak a változó mennyiséggel foglalkozunk. Ennek megfelelő hatást biztosít a hallórendszerben a cochlea ovális ablakon keresztül történő ellenirányú előfeszítése.

A forrás távolsága szerint közeltéri, illetve távoltéri esetről beszélünk. Ez utóbbi esetén, a hullámfront egy sík felülettel közelíthető (lásd 5.2. ábra).

5.2. ábra. Közletér, távoltér.

A hang különböző közegekben különböző sebességgel terjed, levegőben például kb. 344m/s-os sebességgel (c). A terjedés sebessége függ a közvetítő közeg összetételétől, illetve hőmérsékletétől. Periódikus hang hullámhosszán a közvetítő közeg ugyanazon fázisban mozgó részecskéinek legkisebb távolságát értjük. A hullámhossz és a frekvencia között az alábbi arányosság áll fenn:

$$c = f \cdot \lambda \tag{5.2}$$

ahol f a rezgés frekvenciája, λ pedig a hullámhossz.

A hang terjedése a Huygens-, illetve a Huygens-Frensel-elv szerint modellezhető. A hullámfront minden pontja elemi hullámok kiinduló pontja, ezek eredő burkológörbéjének eredménye a hullámterjedés. Inhomogén közegben diffrakció, refrakció, illetve interferencia tapasztalható. Több hullám találkozása esetén interferencia lép fel, azaz az egymással azonos, illetve ellentétes fázisban levő rezgések erősítik, illetve kioltják egymást. A Refrakció, különböző sűrűségű anyagok határfelületén tapasztalható jelenség, mely a hullám terjedési irányának változásával jár. Ennek extrém esete a reflexió, amikor a sűrűségváltozás nagysága miatt a hangenergia egy része visszaverődik, egy kisebb része pedig áthatol az akadályon, valamint legkisebb része a súrlódási veszteség hatására hő formájában felszabadul. A diffrakció, a terjedő hullám útjában levő árnyékoló tárgyak hatása esetén fellépő jelenség, ami hullám elhajlást,illetve az árnyékolt térrészben való hullámképződést eredményez.

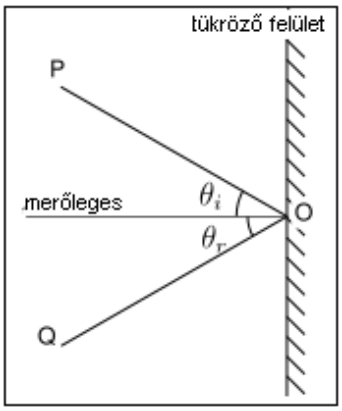
5.2. Akusztikus modellek

A hang terjedése, a fentiek alapján homogén közegben könnyen modellezhető. Véges kiterjedésű inhomogén közeg a gyakorlatban azonban csak süketszobában biztosítható. Egyéb esetekben a zárt teret határoló falak miatt visszaverődések, visszhangok keletkeznek, vagyis az előző fejezetben említett hullámjelenségek lépnek fel [87]. Ezek a hatások az emberi hallgatók számára is érezhető változásokat okoznak a hangtérben, ezért évezredek óta keresettek azok a módszerek, melyekkel a hangtér bizonyos paraméterei - lecsengési idő, hangteljesítmény - az elvárt értékre hangolhatók. Ezek legősibb példái az ókori görög színházak tervezésénél alkalmazott megoldások, melyek segítségével többezer fős arénák hangosítása vált megoldhatóvá.

5.3. ábra. Panorámakép egy ókori görög színházról. Tervezői felismerték a visszaverő felületek hatását, így erősítették fel a színészek hangját úgy, hogy több ezer fő élvezhette az előadást. (forrás: mdoege@compuserve.com http://en.wikipedia.org)

Később méretarányos modellek segítségével igyekeztek a megépítendő színház- és koncerttermek akusztikai viszonyait ellenőrizni, majd a számítástechnika elterjedésével a számítási modellek alkalmazása egyeduralkodóvá vált [88]. A valós körülményeket leghűebben modellező rendszerek a hullámhossz alapján meghatározott felbontásban, végeselem-módszerrel modellezik az akusztikus térben bekövetkező változásokat. Ezek azonban rendkívül számításigényesek, ezért csak speciális esetekben - alacsony frekvencia, kis kiterjedésű akusztikus tér - alkalmazhatóak. Kevésbé precíz, azonban jóval elterjedtebb a hang terjedésének geometriai modell[1] szerinti szimulációja (lásd 5.4. ábra). A hanghullám terjedését a fénysugár terjedésével modellezzük, ami határoló felülettel találkozva a 5.4. ábrán látható módon verődik vissza. Több módszer közölt a megvalósítás részleteire vonatkozóan [89,90], melyek leghatékonyabbjai a sugárkövetéses[2] [91], illetve a nyalábkövetéses[3] [92,93] technikák. Utóbbiakkal hatékonyan és bizonyos feltételek teljesülte esetén pontosan határozhatóak meg komplex terek akusztikai paraméterei. A geometriai hullámterjedési modell csak azokban az esetekben jó közelítés, amikor a határoló felületek mérete lényegesen nagyobb a hang hullámhosszánál, ezért számos kiegészítést dolgoztak ki a szimulációs eredmények javítására [94], ezek tárgyalása azonban jelen dolgozatnak nem képezik tárgyát.

[1] specular reflection method
[2] ray tracing
[3] beam tracing



5.4. ábra. A geometriai hullámterjedési modell.

5.3. Az akusztikus környezet forrás-lokalizációs munkákban használt általános modellje

Az alábbiakban egy, a későbbiekben forrás-lokalizáló módszerek bemutatására használt akusztikus modellt és jelölésrendszert vezetek be. A modell alapja az előző fejezetben tárgyalt geometriai hangterjedési modell, mely egyeduralkodó a forrás-lokalizációval foglalkozó irodalomban.

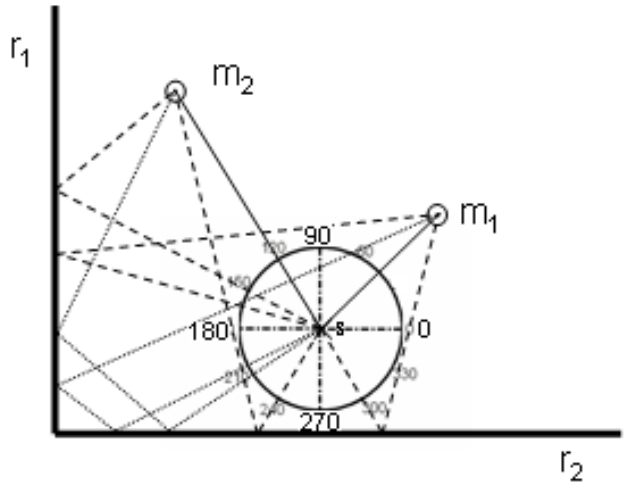
Jelölje tehát s egy pontszerű akusztikus forrás térbeli pozícióját, valamint legyen $s \in C$, ahol C három dimenziós pontok halmaza, melyek a forrás lehetséges térbeli elhelyezkedését reprezentálják. Tegyük fel továbbá, hogy $\xi_s(\varphi, \theta)$[4] kétdimenziós függvény $(0 \leq \xi_s(\varphi, \theta) \leq 1)$ a forrás iránykarakterisztikája, ahol φ jelenti a horizontális, θ pedig a vertikális irányszöget. A rendszer tartalmazzon N darab mikrofont, melyek pozícióit m_i-vel jelölöm $(i = 1 \ldots N, m_i \in C)$. A jelölések egyszerűsítése érdekében tételezzük fel, hogy a mikrofonok azonos típusúak, így $\xi_m(\varphi, \theta)$ kétdimenziós függvény legyen a mikrofon iránykarakterisztikája, $(0 \leq \xi_m(\varphi, \theta) \leq 1)$ ahol φ a horizontális, θ pedig a vertikális irányszöget jelöli. Az akusztikus környezetet visszaverő felületek (r) határolják. Rendeljünk hozzá minden felülethez egy $(0 \leq \beta(r) \leq 1)$ valós számot, mely az r visszaverő felület frekvenciától és beesésiszögtől független abszorpciós koefficiensével egyenlő. Az i. mikrofon és a forrás közötti direkt és a visszaverődések révén létrejövő hangterjedési utakat jelöljük P_i-vel. Az 5.5. ábrán egy kétdimenziós egyszerű példa látható. A bemutatott modellben a mikrofon által rögzített jel az alábbi formában írható:

$$x_i(t) = \sum_{p \in P_i} a(d_p, R_p) \cdot u(t - \tau_p) + \eta_i(t) \tag{5.3}$$

ahol u a forrás által kibocsátott jel időfüggvénye, t az idő, d_p a p út hossza, τ_p a p út megtételézhez szükséges idő, valamint η_i additív, páronként korrelálatlan fehér zaj. A p terjedési út során érintett visszaverő felületek listáját R_p jelöli, míg az α függvény az adott terjedési út során fellépő csillapítás hatását modellezi. E függvény a direkt terjedési út esetén:

$$a(d_p, \{\}) = \frac{1}{d_p} \cdot \xi_s(\varphi_{s,p}, \theta_{s,p}) \cdot \xi_m(\varphi_{m,p}, \theta_{m,p}) \tag{5.4}$$

[4]a dolgozatban a csillapítás értékek minden esetben 0-1 intervallumban értelmezettek



5.5. ábra. Egyszerű kétdimenziós akusztikai rendszer. A direkt terjedési utat folytonos vonal, az egyszeres visszaverődési utat szaggatott, a kétszeres visszaverődési utat pontozott vonal jelöli. Az s forrás körül feltűntettet 0-360 közötti számok a forrás orinentációját megadó irányszögek.

valamely visszaverődés révén létrejövő út esetén pedig:

$$a(d_p, R_p) = \frac{1}{d_p} \cdot \xi_s(\varphi_{s,p}, \theta_{s,p}) \cdot \xi_m(\varphi_{m,p}, \theta_{m,p}) \cdot \prod_{r \in R_p} (1 - \beta(r)) \tag{5.5}$$

alakban írható, ahol $\beta(r)$ az r visszaverő felület abszorpciós koefficiense, $\varphi_{s,p}$ és $\theta_{s,p}$ a p út forrásánál mért horizontális, illetve vertikális irányszög, továbbá $\varphi_{m,p}$ és $\theta_{m,p}$ ugyanezen út i. mikrofonnál mért beesési szögei.

5.4. A forrás-lokalizációval foglalkozó munkák áttekintése

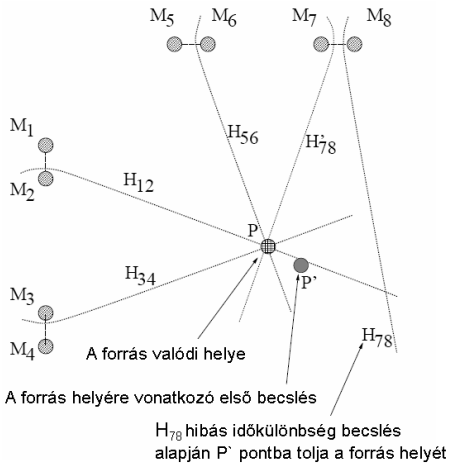
A szenzortömbök által szolgáltatott mérések adatainak alapján történő forráshely meghatározás klasszikus probléma a jelfeldolgozás területén, melynek eredményei egyaránt alkalmazottak az óceanográfia, a radar-technológia és az akusztika területén. Az alábbiakban a hangforrások helyének meghatározását célzó, az elmúlt néhány évtizedben született algoritmusokat tekintem át. A több évtizedes kutatómunka indoka, hogy az eddig elkészült algoritmusok egyike sem ad megnyugtató megoldást, így ma sem rendelkezünk a mindennapi életben alkalmazható, a biológiai rendszerek teljesítményét megközelítő eljárásokkal. A sikertelenség egyik oka, hogy a jeleket általában távol elhelyezett mikrofonokkal rögzítjük, ennélfogva viszonylag rossz jel-zaj viszonnyal bíró, esetleg több jel keverékéből álló elegyből kell kinyerni a forrás helyére utaló információt. Tovább nehezíti a kérdést, hogy a beszéd széles spektrumú periódikus részeket is tartalmazó jel, mely tovább növeli a rendszer szabadságfokainak számát. Végül a mindennapi környezetünket adó zárt terek 0.5-1 másodperc közötti lecsengési ideje, a késletetett jelmásolatok révén nagyban rontja a forrás helyének meghatározására készült eljárások teljesítményét.

A megoldás módja vélhetően több komponensű. Egyrészt kilátástalannak tűnik több forrás jelének keverékét tartalmazó elegyből megállapítani a források helyét. Szükség van az egyes források bizonyos szempontok szerinti előzetes szeparációjára, majd az így azonosított komponensek jeleit felhasználva megkísérelni a forrás-lokalizációt. E stratégia alkalmazására mutattam példát a 4.4. fejezetben. Másrészet kísérleti eredmények igazolják, hogy több mikrofon jelének együttes alkalmazása biztosabb forrás helyre vonatkozó becslést eredményez, ugyanakkor gyakorlati megfontolások okán nem érdemes szélsőségesen nagy mikrfontömböket alkalmazni.

A mikrofontömbök alkalmazásaival foglalkozó, alapműnek tekinthető munkához [95] hasonlóan az algoritmusokat három csoportba sorolva külön alfejezetekben tárgyalom. Egy negyedikben bemutatok egy új, a két legelterjedtebb algoritmuscsalád előnyeit ötvöző módszert. Azokat az eljárásokat, melyek az általam 6. fejezetben bemutatandó megoldásnak az alapjait adják részletesebben tárgyalom.

5.4.1. Érkezési-időkülönbség becslő algoritmusok

Az érkezési-időkülönbség becslő algoritmusok[5] csoportjába tartozó eljárások a forrás helyének meghatározását két lépésben végzik el. Először a mikrofonok jeleinek felhasználásával, mikrofonpáronként igyekeznek meghatározni a jelek közötti - a forrás és a mikrofonok távolságkülönbségéből adódó - időkülönbséget. Az időkülönbség alapján mikrofonpáronként egy hiperbola - háromdimenzióban



5.6. ábra. Az érkezési időkülönbséget becslő algoritmusok működését szemléltető ábra.

hiperboloid - jelölhető ki, mint a forrás lehetséges pozíciói. A mikrofonpárok által kijelölt hiperbolák alapján ideális esetben meghatározható a forrás helye, illetve azokban az esetekben amikor a hiperbolák nem egyetlen pontban metszik egymás, a forrás helyére vonatkozó becslés adható (lásd 5.6. ábra).

A publikált módszerek az imént felsorolt két lépés mindegyikében tartalmaznak különbségeket. Az időkülönbség meghatározására kivétel nélkül valamilyen kereszt-korreláció alapú számítás eredményét felhasználva jutnak, azonban több módszert dolgoztak ki a számítás módjára vonatkozóan [96]. A kereszt-korreláció, definíció szerint

[5] Time Delay of Arrival (TDOA)

$$R_{x_i,x_j}(k) = E[x_i(t) \cdot x_j(t-k)], \tag{5.6}$$

alakban adható meg, ahol E a várható értéket jelöli. Mivel ez a függvény a gyakorlatban nem számítható, az alábbi formulával közelítő becslés adható:

$$c_{x_i,x_j}(k) = \int_{-W}^{W} x_i(t) \cdot x_j(t+k)dt, \tag{5.7}$$

ahol W a korreláció számítására használt ablak hosszának fele. A fenti egyenlet frekvenciatartománybeli alakját általános-kereszt-korrelációs függvénynek[6] nevezzük:

$$c_{x_i,x_j}(k) = \int_{-\infty}^{\infty} \left(G_i(\omega) \cdot X_i(\omega)\right) \left(G_j(\omega) \cdot X_j(\omega)\right)^* e^{j\omega k} d\omega \tag{5.8}$$

ahol * a komplex konjugáltat jelöli, $X_\cdot(\omega)$ az $x_\cdot$ Fourire-transzformáltja, $G_\cdot(\omega)$ pedig tetszőleges szűrő, melytől a gyakorlatban jobban használható kereszt-korrelációs eredményt várunk. Amennyiben ezek a szűrők kontans 1 értékűek, az 5.8. egyenlet megegyezik az 5.7-ben közölt kifejezéssel. Elterjedtebb, a mérsékelten visszhangos körülmények között kiemelkedően jó teljesítményt nyújtó PHAT súlyozás [96] használata, ami az 5.8. egyenlet alábbi átrendezésével:

$$c_{x_i,x_j}(k) = \int_{-\infty}^{\infty} \psi_{i,j}(\omega) \cdot X_i(\omega) \cdot X_j(\omega)^* e^{j\omega k} d\omega \tag{5.9}$$

a $\psi_{i,j}$ súlyfüggvényen keresztül adott, mely

$$\psi_{i,j}(\omega) = \frac{1}{|X_i(\omega) \cdot X_j(\omega)^*|} \tag{5.10}$$

formában írható. A PHAT súlyozás használatával a beérkező jel fehérített változatán végezzük el a számítást, azaz a felhasznált jelekben minden frekvencia azonos súllyal szerepel, tehát a jel periodicitásából adódó korrelációs csúcsok kiküszöbölhetők. Itt érdemes megjegyezni, hogy zajjal terhelt jelek esetén a módszer már kevésbé előnyös hatású, hiszen a jel a zajjal azonos súllyal alakítja a kereszt-korrelációs függvényt, ami fokozott jel-zaj viszonyra való érzékenységet okoz.

Az érkezési időkülönbség a kereszt-korrelációs függvény maximuma alapján

$$\widehat{k_{i,j}} = \max_{k \in D} R_{x_i,x_j}(k) \tag{5.11}$$

formában számítható, ahol a mikrofonok által rögzített jelek között létrejövő legnagyobb időkülönbség (D), a mikrofonok fizikai távolságából adódóan az alábbi formula szerint határozható meg:

$$D = \frac{\|m_i - m_j\|}{c}, \tag{5.12}$$

Az érkezési-időkülönbség meghatározását követően történik meg a becslés térbeli/síkbeli ko-

[6] generalized cross correlation (GCC)

ordinátákká, illetve szögekké való konvertálása. Ezen lépésnél elterjedt a távoltérben, illetve a közeltérben alkalmazható megoldások megkülönböztetése, mivel a hiperbola nehézkes számításából fakadóan más-más egyszerűsítésre nyílik lehetőség. Távoltérben a hiperbola egy, a mikrofonokat összekötő szakasz felezőpontját metsző, adott dőlésszögű egyenessel közelíthető, így a különböző mikrofonpárok eredménye alapján meghatározott egyenesek metszéspontjai könnyedén számíthatóak [97]. Az algoritmusok között találhatók kettő [98–100], illetve három dimenzióban [101] becslést adó megoldások is. Lévén, hogy általában kettőnél több mikrofonpárt használnak a forrás helyének meghatározására, a feladat túlhatározott, azaz egynél több metszéspont alakul ki (lásd 5.6. ábra), melyek közül változatos hibakritériumok szerint történik meg a legvalószínűbb forráshely kiválasztása. A legegyszerűbb módszerek a legnagyobb kereszt-korrelációs csúcshoz tartozó egyenesek, illetve hiperbolák metszete alapján határozzák meg a forrás helyét [98, 99]. A legkorszerűbb eljárások iteratívan választják ki a viszonylag nagy, ám nem feltétlenül a legnagyobb korrelációs értékkel bíró időkülönbséghez tartozó térrészeket [102], illetve léteznek példák a kereszt-korrelációs függvények bizonyos paraméterei (átlag, csúcsossági ráta etc.) alapján kialakított heurisztikus súlyozásra is [103].

Említést érdemelnek még a beszédhang tulajdonságait kiaknázni igyekvő megoldások [100, 104], melyekben a konkurens beszélők hangját egymástól, illetve egyéb zajforrások jelétől különítenek el, így növelve a helymeghatározó módszerek hatékonyságát.

Összefoglalásként elmondható, hogy az ebbe a csoportba sorolható eljárások népszerűsége kis számításigényüknek köszönhető, ugyanakkor nem sikerült általánosan jó megoldást adni az egyes mikrofonpárok különálló becsléseinek akkumulációjára. Ennek eredményeként a viszonylag alacsony számításigény ára a továbbiakban bemutatandó algoritmusokhoz viszonyított szerény teljesítmény.

5.4.2. Nyalábirányítás

Mint az a 5.3. fejezetben látható, a mikrofonok mindegyike rögzíti a sugárzó forrás zajjal és visszaverődésekkel terhelt jelét. Amennyiben a rögzített jeleket a fizikai elhelyezkedésből adódó megfelelő időeltolással összegezzük, a forrásból közvetlen terjedéssel érkező jelek energiája összeadódik, míg a jelhez adott zaj és visszhang energiája az időbeni egyezés hiánya miatt kisebb mértékben növekszik. Ezen gondolatmenet a nyalábirányítási technika[7] klasszikusának a "késleltet és összegez"[8] eljárásnak [105] az alapötlete, mely formálisan:

$$u(t,q) = \sum_{i=1}^{N} x_i(t + \tau_{q,i}) \tag{5.13}$$

alakban írható, ahol $\tau_{q,i}$ a nyalábirányító késleltetés[9], mely a mikrofon tömböt a q pontra $q \in C$ fókuszálja, mely a tömb adott mikrofonjának (i) és a forrás feltételezett helyének (q) távolságából, a hang terjedési sebességének figyelembevételével számítható. Az egyenlet eredményeként kapott $u(t,q)$, a mikrofontömb által rekonstruált forrásjel, azt feltételezve, hogy a forrás a q pontban helyezkedett el. A fentiek alapján a forrás helyét a q helyre vonatkozó jel-energia maximumaként találhatjuk meg:

[7] beamforming
[8] delay and sum beamformer
[9] steering delay

$$\widehat{s} = \max_{q \in C} \left\{ \int_{-W}^{W} u(t,q)^2 dk \right\} \tag{5.14}$$

ahol $\widehat{s}$ jelöli a forrás becsült helyét, W pedig valamely nullánál nagyobb pozitív egész, mely az energia számításhoz felhasznált ablak nagyságát jelöli. A nyalábirányítási technikák a fentiekből következően nem csak a forrás helyének meghatározására használhatóak, hanem egyúttal a forrás jelének kiemelésére is, amit a tömb nyereségnek[10] nevezünk.

Az itt bemutatott nyalábirányítási technikának számtalan továbbfejlesztett változata létezik [106–108], melyek a 5.4.1. fejezetben bemutatottakhoz hasonlóan a jel különbözőképpen szűrt verzióit felhasználva határozzák meg a forrás helyét.

A nyalábirányítási technikák hátránya, hogy a 5.13. és a 5.14. egyenletek eredményét minden lehetséges forráshelyre vonatkozóan ki kell számolni, ami hatalmas számítási kapacitást kíván, ezért komoly energiát fektettek abba, hogy valamilyen módon elkerülhető legyen az összes lehetséges helyre a mikrofontömb válaszának meghatározása. Elterjedt a különféle gradiens keresési eljárások alkalmazása [109], illetve egyéb iteratív, a keresési régió szűkítésén alapuló módszerek [110–112]. Ezek azonban rendkívül érzékenyek a kezdeti feltételekre, valamint megkötéseket tartalmaznak a mikrofon helyére, az akusztikus környezetre, a forrás spektrális tartalmára, esetleg a beszélő mozgására vonatkozóan.

Külön figyelmet érdemelnek az illesztett szűrötömbökkel[11] kombinált nyalábirányítási módszerek [113,114]. Ezek geometriai hullámterjedést feltételezve integrálják az akusztikus környezet hatásait, mivel minden lehetséges forráshelyre meghatározzák a környezet impulzusválasz-függvényét. Az 5.13. egyenletben nemcsak az adott q helyre vonatkozó késleltetett jelek összeadása történik meg, hanem a q helyhez tartozó impulzusválasz-függvény inverzével való szűrés is, ami a visszhang nem kívánt hatását szünteti meg. A módszer sikeresnek bizonyult, jelentős jel-zaj viszony javulást sikerült elérni, ugyanakkor a nagy számításigény miatt, csak speciális és rendkívül drága hardver rendszerekkel vált lehetővé a közel valós idejű alkalmazás [115].

A nyalábirányítási technikák az érkezési-időkülönbség becslő eljárásoknál jobb hatékonysággal, ugyanakkor jócskán megnövekedett számításigényel képesek meghatározni a forrás helyét. A itt tárgyalt módszerek vitathatatlan előnye, hogy segítségükkel kiemelhető a forrás által kibocsátott jel, amit napjaink modern hallókészülékeiben használnak sikerrel. Megjegyzendő azonban, hogy az elérhető jel-zaj viszony javulás az alkalmazott mikrofonok számának nagyjából logaritmusával növekszik, azaz nem minden esetben alkalmazgható megoldás. [102].

5.4.3. Nagyfelbontású spektrális becslők

A nagyfelbontású spektrális becslők csoportjába tartozó módszerek alapvetően a radar technológiában megoldandó problémák megoldására születtek. Az eljárások közös jellemzője a kovariancia mátrix felhasználása alapján, a szenzorok jelei közötti eltérést okozó impulzusválasz-függvény megbecslése autoregresszív modellezéssel, vagy adaptív sajátérték dekompozícióval [116, 117]. Először távoltéri forrás esetén, lineáris geometriájú szenzortömbre alkalmazható megoldás született meg, majd MUSIC algoritmusként elterjedt a tetszőleges geometriájú tömbökre és közeltéri esetre is alkalmazható kiegészítés [118]. Az eljárás alapvetően szűk spektrumú források helyének meghatározására használható, ezért szélessávú jelek esetén több csatorna egyidejű kiszámításával adható

[10] array gain
[11] Matched Filter Array (MFA)

becslés [119–121] erősen növelve a számításigényt. Annak ellenére, hogy ezek a módszerek elterjedtek más szenzortömbökkel kapcsolatos jelfeldolgozási problémák megoldásában, az akusztikus forráshely meghatározás területén nem igazán sikeresek. Az ok, hogy a kovariancia mátrix pontos becslése csak viszonylag hosszú jelszakasz átlagolása alapján lehetséges úgy, hogy ez idő alatt a forrás helye rögzített, jele pedig stacionárius. Beszédhang esetén ezen paraméterek állandóságának, valamint a szükséges átlagolási időnek a biztosítása a gyakorlatban nehézkes.

5.4.4. Akkumulált korrelációs eljárás

Az akkumulált vagy összesített korrelációs eljárás [122] ötvözi az érkezési-időkülönbség becslő módszerek hatékonyságát, a nyalábirányítási technikák robusztusságával. A módszer lényege az egyes mikrofonpárok jeléből számolt kereszt-korrelációs eredmények hatékony összegzése. Az érkezési-időkülönbség becslő eljárásokkal ellentétben a kereszt-korrelációs függvénynek nem csak a maximuma alapján becsüljük a forrás helyét, hanem a mikrofonpáronként számolt kereszt-korrelációs függvényeket egy közös koordináta-rendszerbe vetítjük. A közös koordináta-rendszer lehet a tér egy kitüntetett ponjától mért irány [122], vagy a forrás lehetséges elhelyezkedésének tere [86]. Ez utóbbi eset, formálisan

$$\pounds(l) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} c_{x_i,x_j}(\tau_{i,l} - \tau_{j,l}), \tag{5.15}$$

alakban írható, ahol l a hangforrás egyik lehetséges pozíciója ($l \in C$), $\tau_{i,l}$ és $\tau_{j,l}$ pedig az l pontból a hang terjedéséhez szükséges idő az i, illetve a j mikrofonokhoz, $\pounds(\mathrm{l})$ pedig a hangforrás l helyen való elhelyezkedésének valószínűsége. Az így számolt valószínűségek megegyeznek a nyalábirányítási technikáknál tárgyalt energia értékkel. Különbség mindössze a számítás módjában van, ami a gyakorlatban használt ablakméretek mellett elhanyagolható [123].

5.5. Összefoglalás

A bemutatott algoritmusok kidolgozásába fektetett energia ellenére nem rendelkezünk a gyakorlatban általánosan alkalmazható, megfelelő hatékonysággal bíró megoldásokkal. A módszerek némelyikének számításigénye a ma elérhető számítási kapacitás mellett nem teszi lehetővé a valós idejű alkalmazást, mások nem eléggé robusztusak. Emellett, mint azt a következő fejezetben bizonyítom, az itt bemutatott eljárások visszhangos környezetben a forrás anizotrop iránykarakterisztikája esetén elméleti megfontolások miatt nem adhatnak jó megoldást.

Az akusztikus környezet hatásait integráló forrás-lokalizáló eljárás

6.1. Az akusztikus környezet hatása a kereszt-korrelációs függvényre

Visszhangmentes környezetben - azaz gyakorlatilag kizárólag süketszobában - a kereszt-korrelációs függvény maximuma egyértelműen azonosítja az érkezési-időkülönbséget. Mindennapi környezetünkben azonban a visszhang megbízhatatlanná teszi a kereszt-korrelációs függvény alapján történő forrás meghatározást. Jelen fejezetben a becslés bizonytalanságát okozó, a visszhang hatásaként létrejövő korrelációs csúcsok helyének és méretének becslésére mutatok eljárást, vagyis a visszhang hátrányos hatását használom ki a forrás-lokalizációs probléma hatékonyabb megoldására. Az 5.3. egyenlet 5.7. egyenletbe való behelyettesítésével az alábbi formulát kapjuk:

$$c_{x_i i,x_j}(k) = \int_{t=-W}^{W} \left[\sum_{p\in P_i} a(d_p,R_p)u(t-\tau_p) + \eta_i(t) \right] \cdot \left[\sum_{q\in P_j} a(d_q,R_q)\cdot u(t-\tau_q-k) + \eta_j(t-k) \right] dt$$

$$\begin{aligned} \mathrm{c}_{x_i i,x_j}(k) &= \sum_{p\in P_i}\sum_{q\in P_j} a(d_p,R_p)\cdot a(d_q,R_q)\cdot \left(\int_{t=-W}^{W} u(t-\tau_p)\cdot u(t-\tau_q-k)dt \right) + \\ &+ \sum_{p\in P_i} a(d_p,R_p)\cdot \left(\int_{t=-W}^{W} u(t-\tau_p)\cdot \eta_j(t-k)dt \right) + \sum_{q\in P_j} a(d_q,R_q)\cdot \left(\int_{t=-W}^{W} u(t-\tau_q-k)\cdot \eta_i(t)dt \right) + \\ &+ \int_{t=-W}^{W} \eta_i(t)\cdot\eta_i(t-k)dt, \end{aligned} \tag{6.1}$$

A fenti formula második és harmadik tagja az átlagos zajteljesímény és a jelteljesímény szorzatával egyenlő, míg a negyedik tag az egyes csatornák jeléhez hozzáadott zaj átlag teljesítményének

szorzata. Ezen tagok az alkalmazott modell feltételei szerint - páronként nem korreláló additív zaj, illetve a zajjal nem korreláló jel - konstans tagokkal egyszerűsíthetők. A nem korreláló jelek korrelációjának minimális ingadozását elhanyagoljuk. A fenti meggondolások után, a konstans értékeket elhagyva, valamint az integrál változót $T = t - \tau_p$ -vel helyettesítve a 6.1. egyenlet az alábbi formában írható:

$$c_{x_i,x_j}(k) = \sum_{p \in P_i} \sum_{q \in P_j} a(d_p, R_p) \cdot a(d_q, R_q) \cdot \left(\int_{T=-W-\tau_p}^{W-\tau_p} u(T) \cdot u(T + (\tau_p - \tau_q) - k) dT \right) \tag{6.2}$$

ami nem más, mint az auto-korrelációs függvény különböző súllyal figyelembe vett eltoltjainak összege, azaz:

$$c_{x_i,x_j}(k) = \sum_{p \in P^i} \sum_{q \in P^j} a(d_p, R_p) \cdot a(d_q, R_q) \cdot c_{u,u}(\tau_p - \tau_q - k) \tag{6.3}$$

A későbbi egyszerűsítés érdekében bevezetek egy újabb a fentivel ekvivalens jelölést:

$$c_{x_i,x_j}(k) = \sum_{(p,q) \in P_i \times P_j} a(\tau_p, R_p) \cdot a(\tau_q, R_q) \cdot c_{u,u}(\tau_p - \tau_q - k) \tag{6.4}$$

ahol $\times$ a Descartes-szorzatot jelenti, illetve (p, q) egy rendezett-párt, ahol $p \in P_i$ és $q \in P_j$.

A kereszt-korrelációs függvényt az f és a g visszaverődési utak hatása nélkül

$$c_{x_i,x_j \setminus (f,g)}(k) = \sum_{(p,q) \in P_i \times P_j \setminus (f,g)} a(\tau_p, R_p) \cdot a(\tau_q, R_q) \cdot c_{u,u}(\tau_p - \tau_q - k) \tag{6.5}$$

alakban írható, ahol $f \in P_i$ és $g \in P_j$.

A 6.3. egyenlet kiszámítása a kibocsátott jel (u) ismerete nélkül nem lehetséges, mivel az auto-korrelációs függvény ($c_{u,u}$) nem meghatározható. Másfelől azonban az auto-korrelációs függvény bizonyos tulajdonságainak vizsgálatával fogalmat alkothatunk a kereszt-korrelációs függvény egyes sajátosságairól. Az auto-korrelációs függvény legnagyobb és legmeredekebb csúcsa, lokális maximuma a nulla eltolásnál található (i.e. null csúcs). Az ettől különböző helyeken levő korrelációs csúcsok kisebbek és kevésbé meredekek. Aperiódikus jelek esetén, mint a Dirac delta, az auto-korrelációs függvénynek egyetlen csúcsa van, ezért a kereszt-korrelációs függvény lokális maximumai a 6.4. egyenlet alapján egyértelműen meghatározhatóak, mivel csak a különböző terjedési utak időkülönbségeinek megfelelő helyeken alakulnak ki, az utak csillapításától függő méretű lokális maximumok. Ugyanez igaz más aperiodikus jelek esetén is, ekkor azonban a lokális maximumok alatt azokat a csúcsokat kell értenünk, melyek nagysága szignifikánsan meghaladja a jelek várható értékének szorzatát. Azokban az esetekben, amikor a kibocsátott jel tartalmaz periódikus összetevőket is, mint az a beszédhangok esetén általános, a kereszt-korrelációs függvény visszhang okozta lokális maximumai nem egyértelműen azonosíthatóak a periodicitásból adódó korrelációs csúcsok miatt. Elmondható, hogy azon f és g visszaverődési utak esetén alakul ki lokális maximum, amikor az alábbi két feltétel teljesül:

$$a(\tau_f, R_f) \cdot a(\tau_g, R_g) \cdot c_{u,u}(0)'_+ > c_{x_i,x_j \setminus (f,g)}(\tau_f - \tau_g)'_+$$
$$a(\tau_f, R_f) \cdot a(\tau_g, R_g) \cdot c_{u,u}(0)'_- > c_{x_i,x_j \setminus (f,g)}(\tau_f - \tau_g)'_- \quad (6.6)$$

ahol a $c_{u,u}(0)'_-$, illetve a $c_{u,u}(0)'_+$ tagok az auto-korrelációs függvény nulla helyén a bal, illetve a jobb oldali deriváltakat jelölik, míg a $c_{x_i,x_j \setminus (f,g)}(\tau_f - \tau_g)'_-$, illetve a $c_{x_i,x_j \setminus (f,g)}(\tau_f - \tau_g)'_+$ kifejezés a kereszt-korrelációs függvény f és g visszaverődési utak hatása nélküli alakjának $(\tau_f - \tau_g)$ helyen számolt bal, illetve jobb oldali deriváltjái. A fenti feltételek teljesülése a kibocsátott jel ismerete nélkül pontosan nem meghatározható, azonban elmondható, hogy a kereszt-korrelációs függvény $(\tau_f - \tau_g)$ helyén lokális maximum kialakulása valószínűsíthető, amennyiben a

$$a(\tau_f, R_f) \cdot a(\tau_g, R_g) \cdot c_{u,u}(0) >> c_{u,u}(h)(h \neq 0) \quad (6.7)$$

feltétel teljesül, azaz amennyiben az adott visszaverődési út csillapítása kicsi és a kibocsátott jel auto-korrelációs függvényének nullától különböző helyeken felvett értéke közel zérus, azaz a jel nem periódikus.

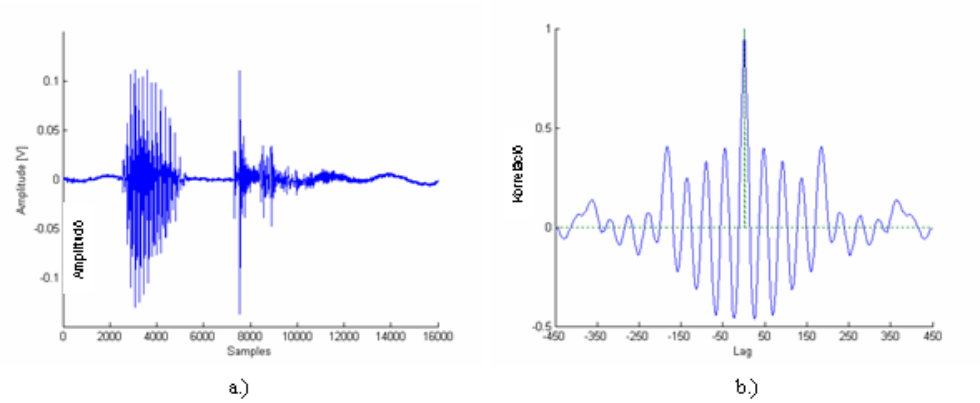
A 5.4.1. fejezetben már említett PHAT súlyozás használatával [96] a bejövő jel aperiodikussá tehető, tehát a második feltétel teljesíthető. A fentiek figyelembevételével definiálható a kereszt-korrelációs függvény lokális maximum helyeit jósló függvény, mely:

$$p_{x_i,x_j}(k) = \sum_{p \in P_i} \sum_{q \in P_j} a(\tau_p, R_p) \cdot a(\tau_q, R_q) \cdot \delta(\tau_p - \tau_q - k) \quad (6.8)$$

alakban írható, ahol a $\delta(\tau_p - \tau_q - k)$ a Dirac delta függvény $(\tau_p - \tau_q)$-val való eltoltjának k helyen felvett értékét jelenti. A fenti függvény természetesen nem jósolja meg a kereszt-korrelációs függvény minden egyes maxium helyét. Lehetnek további lokális maximumok a beérkező jel periodicitásából adódóan, emellett az erősen csillapított visszaverődési utak nem szükségszerűen okozzák lokális maximum kialakulását. Éppen ezért, az imént definiált a kereszt-korrelációs függvény lokális maximumait jósoló függvény $(p_{x_1,x_2}(k))$, a kereszt-korrelációs függvény adott helyen levő lokális maximumának valószínűsége szerint is értelmezhető, jóllehet ebben az esetben a „valószínűség" nem a szigorú matematikai értelemben vett valószínűséget jelenti. A 6.2. ábra felső részén a 6.1. ábrán látható jel kereszt-korrelációs függvényét ábrázoltam az 5.5. ábrán látható akusztikus környezetet és elrendezést feltételezve. A modellben a jelek rögzítésére használt mikrofonok omnidirekcionálisak, a forrás pedig izotróp iránykarakterisztikájú. A 6.2. ábrán megfigyelhető, hogy azokon a helyeken, ahol a predikció lokális maximumot jelöl valóban lokális maximum alakul ki. A PHAT súlyozással számolt kereszt-korrelációs függvény esetén a lokális maximumok jósoltakkal való egybeesése még szembetűnőbb.

A 6.2. ábrán a kereszt-korrelációs függvényen négyzetekkel jelöljük a terjedési út-párok időkülönbségének megfelelő helyeket. Ezeket a helyeket két számmal azonosítjuk. Az első szám az m_1, míg a második az m_2 mikrofont elérő terjedési utat kódolja, úgymint: 1 - direkt terjedési út (folytonos vonallal jelzett a 5.5. ábrán); 2, 3 - egyszeres visszaverődési út (szaggatott vonal); 4 - kétszeres visszaverődési út (pontozott vonal). Az ábrán jól látható, hogy ebben az esetben a forrásból a mikrofonokat közvetlenül elérő (1-1) terjedési utak csillapítása a legkisebb, így ezen a helyen található a kereszt-korrelációs függvény maximuma.

A kereszt-korrelációs függvény terjedési utak okozta lokális maximum helyei az akusztikus környezettől függenek, éppen ezért - feltéve, hogy a visszaverő felületek elhelyezkedését és paramétereit



6.1. ábra. A forrás által kibocsátott jel, egy férfi beszélő által kiejtett, süketszobában rögzített 'ok' szó (a.), és ennek auto-korrelációs függvénye (b.).

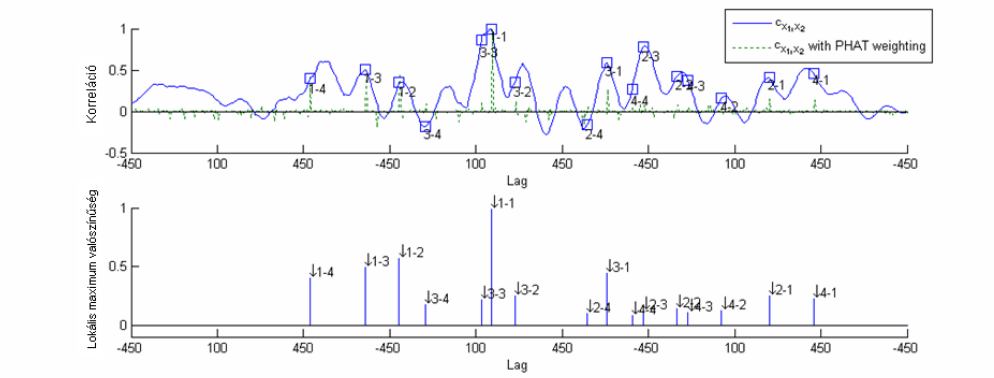
változatlannak tekintjük - a hangforrás helyétől függően más-más lokális maximum jósló függvények jönnek létre. A továbbiakban a $p_{s,x_i,x_j}(k)$ a $p_{x_i,x_j}(k)$ függvényt jelöli abban az esetben, amikor a sugárzó hangforrás az s pontban van.

6.1.1. Anizotrop források hatása

Az eddig közölt forrás-lokalizációval foglalkozó munkák egyikében sem vizsgáltak az anizotrop források esetén fellépő effektusok, jóllehet, mint arra a későbbiekben rámutatok, bizonyos esetekben alapvető fontosságúak lehetnek. Napjainkban az érkezési-időkülönbség becslő algoritmusok gyenge teljesítménye általánosan azzal magyarázott, hogy a visszhang által okozott téves korrelációs csúcsok rontják a becslés megbízhatóságát, noha a valós érkezési-időkülönbséget jelző csúcsnál csak abban az esetben alakulhat ki nagyobb korrelációs csúcs, ha több visszaverődési út hatása összegződik. A forrás, illetve a mikrofonok iránykarakterisztikájának figyelembevételével a fenti magyarázathoz fontos kiegészítéseket lehet fűzni. Többé nem szükséges feltétel a visszaverődési utak által okozott korrelációs csúcsok időbeni egybeesése, hiszen a direkt utak időkülönbségét jelző lokális maximumnál nagyobbat eredményezhet egy kevésbé csillapított visszaverődési út. Beszélők helyének meghatározása esetén a fej, illetve a száj együttes hatása által létrehozott iránykarakterisztikát kell figyelembe vennünk, - természetesen az esetleges nem omnidirekcionális mikrofonok karakterisztikája mellett - ami több dB-es csillapítási különbséget okozhat a hang frekvenciájától, illetve a terjedési út eredőjénél mért vertikális és horizontális irányszögtől függően [124] (lásd 6.3. ábra). Lévén, hogy a dolgozatban alkalmazott modell a frekvenciától független, a száj iránykarakterisztikáját frekvenciafüggetlen átlagként veszem figyelembe. Ezzel az egyszerűsítéssel élve kijelenthető, hogy azokban az esetekben, amikor a

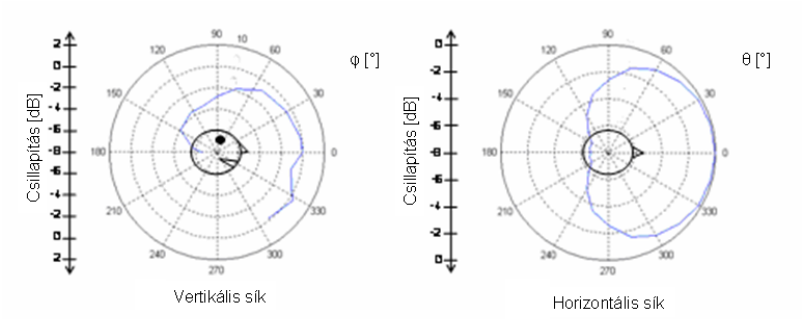
$$\alpha(\tau_d, \{\}) < \alpha(\tau_r, R_r) \tag{6.9}$$

feltétel teljesül, - amelyben az r és d indexek tetszőleges visszaverődési, illetve direkt utat jelölnek - a kereszt-korrelációs függvény maximuma nem a valós érkezési-időkülönbséget azonosítja. Meg-



6.2. ábra. A kereszt-korrelációs függvény (felső ábra) és ennek jósolt lokális maximum helyei (alsó ábra).

jegyzendő, hogy az $\alpha(.,.)$ függvény értékkészlete a 0-1 intervallumra korlátozódik, tehát a nagyobb csillapítás kisebb α értékkel jár.

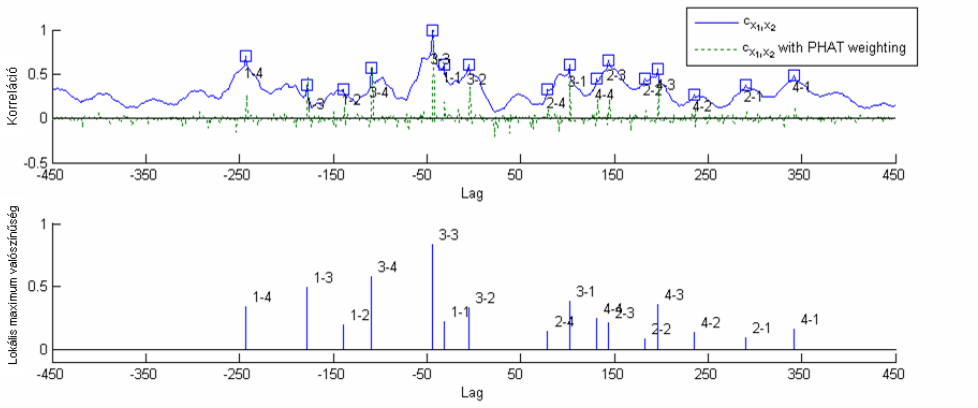


6.3. ábra. Átlagos beszélő szájának iránykarakterisztikája. Az ábrázolt csillapítás értékek 160Hz-től 8kHz-ig terjedő, harmad oktávonként mért eredmények átlaga. (A [124]-ban közölt adatok alapján.)

A forrás iránykarakterisztika hatásának szemléltetése érdekében az 5.5. ábrán látható akusztikus környezetbe helyeztem egy, a 6.3. ábrán látható iránykarakterisztikájú beszélőt. A beszélő irányát az 5.5. ábrán feltüntetettnek megfelelően 270^o-ra választottam. A modell által szolgáltatott jelek segítségével számolt kereszt-korrelációs függvények a 6.4. ábrán láthatóak.

Mint az a 6.4. ábrán látható, a kereszt-korrelációs függvény legnagyobb csúcsa a (3-3) kóddal jelölt helyen van, azaz két kevéssé csillapított visszaverődési út okozza a legnagyobb korrelációs csúcsot, tehát hibás helymeghatározás történik a hagyományos érkezési-időkülönbség becslő algoritmusok [86, 96–101, 122, 123] esetén.

A helyes érkezési-időkülönbség megtalálásához tehát, anizotrop forrás esetén, figyelembe kell vennünk a forrás irányát is, ezért a lokális maximum becslő függvény definíciójánál elengedhetetlen a forrás irányának rögzítése. A továbbiakban $p_{s,\varphi,\theta,x_i,x_j}(k)$, az s pontban elhelyezett for-



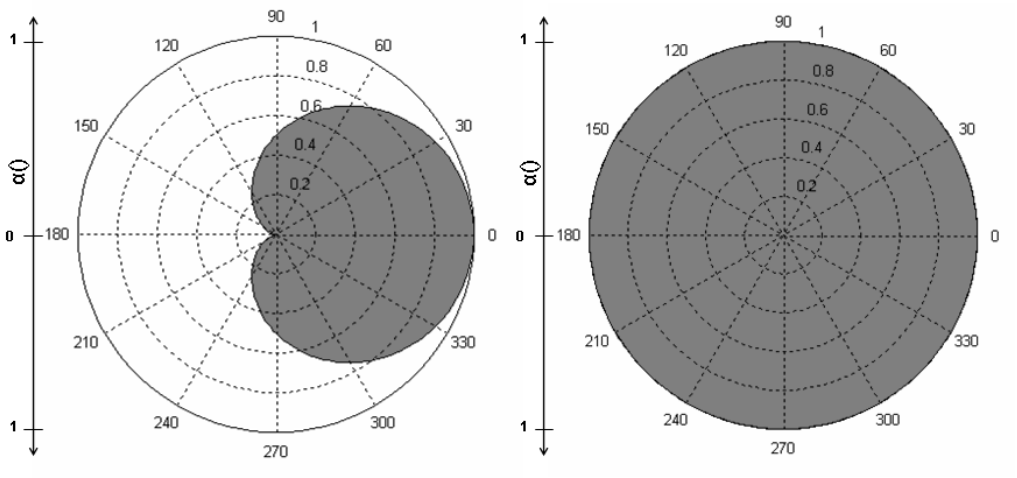
6.4. ábra. Az emberi beszélő iránykarakterisztikájának hatása a hagyományos érkezési-időkülöbség becslő algoritmusokra. A helyes érkezési-időkülönbség az (1-1)-gyel jelölt eltolásnál van, noha a kereszt-korrelációs függvény maximuma a (3-3)-mal jelölt helyen található.

rás φ és θ horizontális és vertikális irányszöge esetén, az i és a j mikrofonok által rögzített jelekből számolt kereszt-korrelációs függvény lokális maximum becslő függvényét jelöli. A kereszt-korrelációs függvény lokális maximumai minden mikrofonpár és minden lehetséges akusztikus konfiguráció esetén meghatározandóak. Amennyiben az akusztikus környezetet változatlannak tekintjük ez $\binom{N}{2} \cdot |C_A|$ darab lokális maximum becslő függvényt jelent, ahol N a mikrofonok száma, $|C_A|$ pedig a lehetséges forrás-konfigurációkat tartalmazó halmaz elemszámát jelöli. A C_A elemei olyan rendezett-hármasok, melyek általánosan (s, φ, θ) alakban írhatóak fel, ahol s a forrás helyét jelöli, míg φ és θ a forrás horizontális, illetve vertikális irányának szöge. Magától értetődő módon, izotróp források esetén nincs szükség a különböző irányok megkülönböztetésére, ezért ebben az esetben $|C_A| = |C|$.

A közölt példák mindegyikében izotróp mikrofonok használatát feltételeztem, ugyanakkor fontos megjegyezni a modell lehetővé teszi tetszőleges $\xi_m(\varphi, \theta)$-val jelzett iránykarakterisztika használatát. Anizotróp mikrofon karakterisztika esetén, mint például a 6.5. ábra jobb oldalán ábrázolt félcardioid karakterisztika, megnő azon esetek száma, amikor a 6.9. feltétel teljesül, mégpedig akkor amikor a beszélő a mikrofon egy kevésbé kiemelt térrészében tartózkódik.

6.2. Az akusztikus környezet hatásának akkumulációja

A kereszt-korrelációs függvény lokális maximumainak mikrofonpáronkénti becsléseinek alkalmas összegzése alapvető fontosságú az algoritmus robusztus és hatékony működésének szempontjából. Az 5.4.4. fejezetben bemutatott eljárás alkalmas e feladat megoldására, azonban esetünkben nem a kereszt-korrelációs függvény közös koordináta-rendszerbe vetítéséről, mint inkább a lokális maximum becslések közös koordináta-rendszerbe vetítéséről van szó, ezért az 5.15. egyenlet analógiájára definiálom a



6.5. ábra. Mikrofon iránykarakterisztikák. A bal oldalon egy fél-cardioid, míg a jobb oldalon egy izotróp karakterisztika látható.

$$\mathrm{p}^{RM}_{s,\varphi,\theta}(\mathrm{l}) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} p_{s,\varphi,\theta,x_i,x_j}(\tau_{i,l} - \tau_{j,l}) \tag{6.10}$$

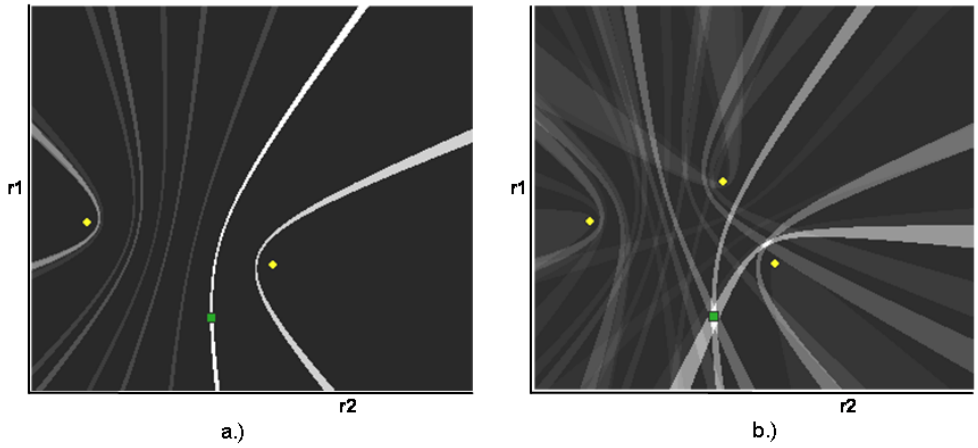
kifejezést, ahol $\mathrm{p}^{\mathrm{RM}}_{s,\varphi,\theta}(\mathrm{l})$ az $l \in C$ pontban összesített jósolt lokális maximumok értékét jelöli az $(s,\varphi,\theta) \in C_A$ akusztikus konfiguráció esetén. Mivel a lokális maximum kialakulásának esélye függ a késleltetett visszaverődések csillapításától, $\mathrm{p}^{\mathrm{RM}}_{s,\varphi,\theta}(\mathrm{l})$ nem más, mint a lokális maximum kialakulásának valószínűsége az l pontban. A $\mathrm{p}^{\mathrm{RM}}_{s,\varphi,\theta}(\mathrm{l})$ minden lehetséges forráshelyre való kiszámításával az úgynevezett becsült visszhanghatás-térképet kapjuk, melyet $\mathrm{p}^{\mathrm{RM}}_{s,\varphi,\theta}$ jelöl. A 6.6. ábra bal oldalán az 5.5. ábrán látható elrendezés becsült visszhanghatás-térképe látható, míg az ábra jobb oldala ugyanezen elrendezés becsült visszhanghatás-térképe három mikrofon esetén. Az ábrán jelzett esetekben mind a forrás, mind a mikrofonok anizotrop iránykarakterisztikájúak.

A becsült visszhanghatás-térképek legjellegzetesebb pontjai a lokális maximumok, ezért ezek egy részhalmazára bevezetem a

$$\widehat{\widehat{p^{RM}_{s,\varphi,\theta}}} = \left\{ m \in \widehat{p^{RM}_{s,\varphi,\theta}} | p^{RM}_{s,\varphi,\theta}(m) > T_r \cdot \max_{c \in C} \{ p^{RM}_{s,\varphi,\theta}(c) \} \right\} \tag{6.11}$$

jelölést, ahol a T_r paraméter a legkisebb figyelembe vett visszhanghatás értékét[1] adja meg, míg $\widehat{p^{RM}_{s,\varphi,\theta}}$ a térkép összes lokális maximumának a jele. A továbbiak könnyebb követhetősége érdekében megjegyzem, hogy egyszeres kalap $\widehat{(\cdot)}$ jelölést használok tetszőleges térkép lokális maximum helyeinek jelzésére, míg dupla kalappal $\widehat{\widehat{(\cdot)}}$ a lokális maximumok egy adott határt (T_r) meghaladó részhalmazát jelölöm.

[1] A *figyelembe vett visszhanghatás szint*, tulajdonképpen a térkép legnagyobb lokális maximumához viszonyított arányt jelöli.



6.6. ábra. Becsült visszhanghatás-térképek. A zöld négyzet a forrás, míg a sárga pontok a mikrofonok helyét jelzi. Az ábra a.) részén az 5.5. ábrán látható elrendezés becsült visszhanghatás-térképe látható, míg a b.) térkép ugyanezen elrendezés becsült visszhanghatás-térképe három mikrofon esetén.

6.3. Az inverz probléma megoldása

A forrás-lokalizáció problémájának megoldása során az algoritmus bemenetei a mikrofonok (<10) által rögzített jelek, melyekből mikrofonpáronként kereszt-korreláció számítható. A kereszt-korrelációs függvényt az 5.15. egyenletben leírt módon közös koordináta-rendszerbe vetítem. Amennyiben az összes lehetséges forráshelyre kiszámítjuk a forrás elhelyezkedésének valószínűségét, az úgynevezett összesített korrelációs térképet ($\mathcal{L}$) kapjuk. Birchfield [86] a legnagyobb valószínűséggel bíró forráshelyet választja ki a forrás hipotetikus helyeként. Jelen munkában az adott helyen található forrás által létrehozott visszhang tulajdonságait is figyelembevéve hozunk döntést. Mint azt korábban bemutattam a visszhang lokális csúcsokat eredményez a kereszt-korrelációs függvényen, melyek kiemelésére a PHAT súlyozás használható. A kereszt-korrelációs függvényeket felhasználva létrehozható az összesített korrelációs térkép $\mathcal{L}$, melynek lokális maximum helyeinek megtalálásával a rögzített jelek visszhang okozta hatása vizsgálható. Ennek megfelelően a

$$\widehat{\widehat{\mathcal{L}}} = \left\{ m \in \widehat{\mathcal{L}} | \mathcal{L}(m) > T_r \cdot \mathcal{L}_{\max} \right\} \tag{6.12}$$

formulával definiálom a megfigyelés alapján azonosított visszhanghatásokat, ahol T_r továbbra is a figyelembe vett legkisebb visszhanghatás mértékét, $\widehat{\mathcal{L}}$ pedig az összesített korrelációs térkép lokális maximum helyeit jelöli, ahol $\mathcal{L}_{\max} = \max_{l \in C}\{\mathcal{L}(l)\}$-val egyenlő.

6.3.1. A legjobban illeszkedő tárolt konfiguráció kiválasztása

Az előző fejezetekben megmutattam hogyan készíthető becslés az akusztikus környezet, illetve a forrás helyének figyelembevételével a kereszt-korrelációs függvény lokális maximumaira, valamint módszert adtam a kereszt-korrelációs függvényből a visszhang hatásainak kinyerésére. Jelen feje-

zetben a megfigyeléshez legjobban illeszkedő becslés kiválasztását ismertetem.
Az algoritmus első lépése a szóba jöhető akusztikus konfigurációk - meglehetősen durva szempont szerinti előzetes - kiválasztása, az úgynevezett lehetséges konfigurációk halmazának (f_C) létrehozása. A lehetséges konfigurációk halmazába (f_C) azok a konfigurációk ($f_C = \{(z,\varphi,\theta) \in A_C\} \subset A_C$) tartoznak, melyek becsült visszhanghatás-térképének maximum helyén ($m \in C, p^{RM}_{z,\varphi,\theta}(m) = \max_{l\in C}\{p^{RM}_{z,\varphi,\theta}(l)\}$) a megfigyelések szerint is közel maximális érték található ($\pounds_{\max} \cdot T_c < \pounds(m)$). A következőkben ezen lehetséges konfigurációk halmazából (f_C) választom ki a legvalószínűbb forráshelyet. Emlékeztetőül jegyzem meg, hogy mind a becsült visszhanghatás-térkép lokális maximumai, melyek minden egyes akusztikus konfiguráció esetére meghatározottak, mind a megfigyelések alapján készített visszhang hatás térképek a megfigyelt tér pontjait tartalmazzák. A pontokhoz minden esetben egy érték, a térkép adott helyén levő lokális maximum mérete rendelhető. A becsült lokális maximum pontok száma egy-egy akusztikus konfiguráció esetén más-más lehet, mitöbb, a megfigyelések alapján létrehozott visszhanghatást jellemző lokális maximum pontok száma is változhat a jelekhez hozzáadódott zaj miatt. A feladat tehát változó elemszámú ponthalmazok hasonlóságának meghatározása, amire globális paraméterek, - ilyen például a súlypont - segítségével következtethetünk. A fenti megfontolásokat figyelembe véve, az alábbi hasonlósági mértéket definiálom a megfigyelés alapján készített ponthalmaz ($\widehat{\widehat{\pounds}}$) és az lehetséges konfigurációk halmazában (f_C) levő ponthalmazok ($\widehat{\widehat{p^{RM}_{z,\varphi,\theta}}}$) között:

$$\begin{aligned} D(z,\varphi,\theta) = ||P_{cg}(\widehat{\widehat{p^{RM}_{z,\varphi,\theta}}}) - P_{cg}(\widehat{\widehat{\pounds}})|| + \\ + \ ||P_{icg}(\widehat{\widehat{p^{RM}_{z,\varphi,\theta}}}) - P_{icg}(\widehat{\widehat{\pounds}})|| \end{aligned} \tag{6.13}$$

A fenti kifejezés első tagja a (z,φ,θ) konfiguráció becsült visszhanghatás-térképének lokális maximum helyeinek súlypontjának és a megfigyelés alapján létrehozott visszhanghatás-térkép lokális maximum helyeinek súlypontjainak távolságát jelenti. Tetszőleges $M \in \left\{\widehat{\widehat{p^{RM}_{z,\varphi,\theta}}} | (z,\varphi,\theta) \in f_C\right\} \cup \left\{\widehat{\widehat{\pounds}}\right\}$ ponthalmaz súlypontjának kiszámítása a

$$P_{cg}(M) = \frac{\sum\limits_{m\in M}(M(m)\cdot T_{TDOA}(m))}{\sum\limits_{m\in M} M(m)} \tag{6.14}$$

kifejezés szerint történik, ahol $M(m)$ az m pont M térképen felvett értékével egyenlő, $T_{TDOA}(m)$ egy $\begin{pmatrix} N \\ 2 \end{pmatrix}$ dimenziós vektor, mely az m pont helyét jelöli az érkezési-időkülönbségek terében ($\mathbb{S}_{TDOA}$). $\left(T_{TDOA}(m) \in \mathbb{S}_{TDOA} \subset \mathbb{R}^{\begin{pmatrix} N \\ 2 \end{pmatrix}}\right)$ $T_{TDOA}(.)$ egy C-ből $\mathbb{S}_{TDOA}$-ba vetítő transzformáció:

$$T_{TDOA}(m) = (\chi_1, \chi_2, ..., \chi_{\begin{pmatrix} N \\ 2 \end{pmatrix}})^T \tag{6.15}$$

ahol $\chi_k = \tau_{i,m} - \tau_{j,m}$ $\left(k = 1... \begin{pmatrix} N \\ 2 \end{pmatrix}\right)$ az érkezési-időkülönbségek terének k-adik koordinátája, $\tau_{i,m}$ és $\tau_{j,m}$ a hang m pontból i, illetve j mikrofonokig tartó útjához szükséges idő. A mikrofonindexek kiválasztása rendezett-párok formájában történik, ahol (i, j) a mikrofonindexek lehetséges kombinációiból képzett lista k. tagja. A $P_{cg}(M)$ eredménye az M ponthalmaz $\mathbb{S}_{TDOA}$-beli súlypontja.

A 6.13. egyenlet második tagja az úgynevezett inverz súlypontok távolsága, mely inverz súlypont az alábbi módon számítható:

$$P_{icg}(M) = \frac{\sum\limits_{m \in M} [(M_{\max} - M(m)) \cdot T_{TDOA}(m)]}{\sum\limits_{m \in M} (M_{\max} - M(m))} \tag{6.16}$$

ahol M_{max} az M térkép maximuma ($M_{max} = \max_{l \in C}\{M(l)\}$).
A 6.13. egyenletben a $||.||$ adott $\mathbb{S}_{TDOA}$-beli vektor hosszát jelöli, ami ebben az esetben nem más, mint egy tárolt becslés és a megfigyelés távolsága:

$$\|v_{TDOA}\| = \sum_{k=1}^{\begin{pmatrix} N \\ 2 \end{pmatrix}} \sqrt{\nu_k^2} \tag{6.17}$$

ahol ν_k a $v_{TDOA} \in \mathbb{S}_{TDOA}$ vektor k. koordinátája.
A hipotetikus forráshelyet a megfigyelésekhez legjobban illeszkedő ponthalmazhoz tartozó konfiguráció adja, mely az alábbi módon választható

$$\widehat{s} = \min_{(z,\varphi,\theta) \in f_C} \{D(z, \varphi, \theta)\} \tag{6.18}$$

A fenti módszerrel történő hangforrás-lokalizációt a későbbiekben Anizotrop ForrásHely Meghatározó (AFHM) algoritmusnak nevezem.

6.4. A diszkretizáció

A fejezet eddig felírt formulái és megállapításai folytonos idő változót feltételezve, illetve végtelen finomságú rács metszéspontjai mentén elhelyezett lehetséges forráshelyek esetére vonatkoztak, mely feltételek a gyakorlatban nem biztosíthatóak. Feltételezve, hogy minden késleltetés ($\tau_{i,c}$, $c \in C$, $i = 1 \dots N$) felbontható a mintavételi idő egész számú többszörösére, a Nyquist-tétel segítségével a folytonos időváltozók diszkrét ekvivalenseikkel helyettesíthetők. Az összesített korrelációs térkép ($\pounds$) térbeli felbontásának kérdése a nyalábirányítási technikáknál jól ismert problémára vezet, melynek lényege a felbontás durvaságából fakadó időbeni pontatlanság okozta hibás forráshely meghatározás[2] [110]. A nyalábirányítási technikák energiatérképe a nyalábirányító rendszer adott helyre vonatkozó kimeneti energiája, mely a forrás valós helyén maximális értékű. Ezen csúcs energia térképen való kiterjedése a forrás által kisugárzott frekvencia nagyságával fordított arányban csökken. A közölt munkában [110] becslést adtak az említett maximum kiterjedésére, valamint megfogalmazták, hogy amennyiben a kisugárzott jel legnagyobb frekvenciájához tartozó hullámhossz ötödénél

[2] problem of time delay imprecision or misalignment of beamformers

kisebb hibával közelítjük a forrás helyét[3], koherens energia többletet állapíthatunk meg a forrás valós helyének megfelelő pozícióban. Mivel az összesített korrelációs térkép lényegében azonos a nyalábiráyítási technikák energia térképével [123], a fenti eredmény esetünkben is alkalmazható, azaz a maximális megengedhető térbeli felbontás alapján meghatározható a lokalizációhoz felhasznált legnagyobb frekvencia. Ugyanezen elgondolás alapján megoldható a becsült lokális maximum függvények közös koordináta-rendszerbe vetítése, mint az a 6.10. kifejezésben látható, azonban a $p_{x_i,x_j}(k)$ kifejezés újradefiniálására van szükség, a következő módon:

$$p_{x_i,x_j}(k) = \sum_{p\in P_i}\sum_{q\in P_j} a(\tau_p, R_p)\cdot a(\tau_q, R_q)\cdot \Pi(\tau_p - \tau_q - k) \tag{6.19}$$

ahol $\Pi(\tau_p - \tau_q - k)$ jelenti a Dirac delta felső frekvenciáktól szűrt és $(\tau_p - \tau_q)$-val eltolt verziójának k-ban felvett értékét. A szűrő határfrekvenciáját a [110]-ben közöltek alapján választottam. A lokális maximum becslő függvények imént közölt változatát használva a becsült visszhanghatás-térkép ($p^{RM}_{s,\varphi,\theta}$) tetszőleges sűrűségű rács esetén megadható.

6.5. A módszer teljesítményének vizsgálata

6.5.1. A teszt környezet

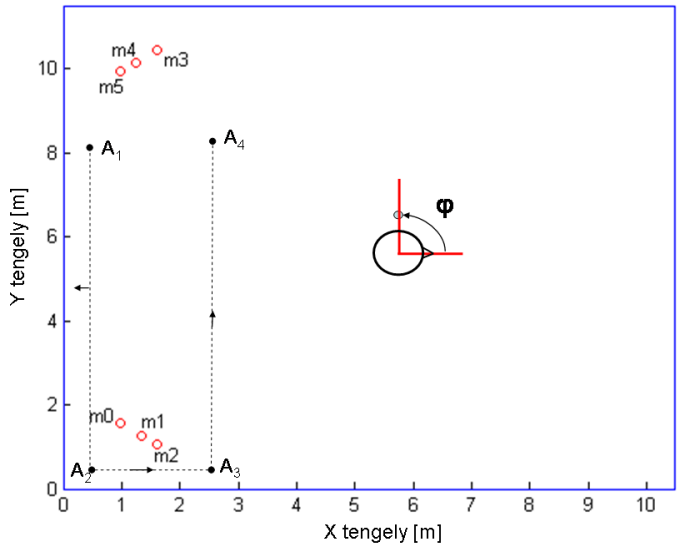
A forrás-lokalizációval foglalkozó algoritmusok teljesítményét rendszerint visszhangos, illetve zajos körülményekkel szembeni robusztusságként értelmezik. Ezek ellenőrzésére elkészítettem a Pázmány Egyetem Práter utcai épületében található előadóterem akusztikus modelljét a CATT[4] [125] szoftver segítségével. Az előadóterem háromdimenziós modelljében (6.7. ábra) 1.7m magassan, egy úgynevezett forrás-síkot definiáltam, mely sík, a forrás lehetséges pozícióit jelöli, annak feltételezésével, hogy az átlagos beszélő szája ebbe a magasságba esik. Ez az esetek többségében megfelelő pontosságot eredményez az ettől különböző magasságú beszélők esetén is, mivel az alkalmazott nagy mikrofontávolságok miatt a magasságkülönbségből adódó időkülönbség eltérés az esetek többségében nem haladja meg azt a szintet, mely az alkalmazott mintavételi frekvencia mellett kimutatható lenne. Az emberi hang a beszélő nemétől függően az 500Hz-től 700Hz-ig terjedő tartományban hordozza a legtöbb energiát, ezért a modell validációjához 700Hz-et választottam, mint a lokalizációhoz felhasznált legmagasabb frekvencia. A legmagasabb frekvencia megválasztása alapján a [110]-ben publikált eredmények szerint meghatároztam a lehetséges forrás pontok (C) felbontását, mely egy 0.1m sűrűségű négyzetes rácsot eredményezett a forrás-síkon.

A becsült lokális maximum függvények elkészítéséhez elengedhetetlen az összes lehetséges forráshelyre vonatkozóan az egyes visszaverődési utak, azaz az akusztikus környezet impulzusválaszának ismerete. A lehetséges forráshelyek nagy száma meglehetősen problémássá, de mindenképpen időigényessé teszi a szoba akusztikus paramétereinek kísérletekkel történő meghatározását, ezért tesznek jó szolgálatot a napjainkban már kereskedelmi forgalomban kapható akusztikus modellező szoftverek [125, 126], melyek komplex geometriájú terek impulzusválasz-függvényének meghatározására is alkalmasak. Jelen munkában a már említett CATT programot használtam a terem átviteli-függvényének meghatározására. A lehetséges forráshelyek halmazának (C) minden pontjában meghatároztam a 0, 90, 180 és 270^o-os beszélő irányhoz tartozó impulzusválasz-függvényt. Mint az a 6.3. ábrán látható, a száj iránykarakterisztikája a vertikális síkban a $\pm 60^o$-os, gyakorlati

[3] $\lambda/5$ imprecision heuristic
[4] Computer Aided Theater Technique

6.7. ábra. A modellezett és a valós környezet képe.

szempontból kitüntett jelentőségű tartományban közel homogén, ezért feltételeztem, hogy a modellezett beszélő forrás-síkkal bezárt vertikális szögének 0^o-ban való rögzítése eltérő irányok esetén is jó közelítést ad. Kísérleteimben a száj iránykarakterisztikáját a [124]-ban közölt, 1kHz alatti csillapítás értékeket átlagolva határoztam meg, mely jó közelítéssel használható tetszőleges nemű beszélő iránykarakterisztikájának modellezésére [124]. A kísérleti környezetben használt mikrofonok helye és a beszélő irányszögének értelmezése a 6.8. ábrán látható.



6.8. ábra. A teremben elhelyezett mikrofonok helye, valamint a kísérletekben modellezett előadó útja (pontozott vonal).

A fenti módszer 53891 különféle akusztikus konfigurációt és 323346 impulzusválasz-függvényt eredményezett. A becsült visszhanghatás-térképeket az öt legerősebb visszaverődést figyelembe véve készítettem el, a 6.19. egyenletben leírtaknak megfelelően, 25kHz-es mintavételi frekvenciát felté-

telezve. Az egyes térképek lokális maximumait ($\widehat{p^{RM}}$, $\widehat{\mathcal{L}}$) 1077 egymást követően végrehajtott gradiens keresés eredményeként határoztam meg. A keresések kezdőpontjaiként a térképen egyenlő távolságban elhelyezett 1077 pont szolgált. Az előzetes számítások időigénye a **??**. táblázatban látható. A lehetséges konfigurációk halmazát (f_C) az összesített korrelációs térkép maximumának

A CATT programmal az összes lehetséges akusztikus konfigurációban meghatározni a mikrofonok impulzusválasz-függvényét.	~12 h
A CATT kimenetét konvertáló C++ program futási ideje.	~ 2 h
Az visszhanghatás-térképek elkészítése és a jellemző pontok megkeresése.	~6 h

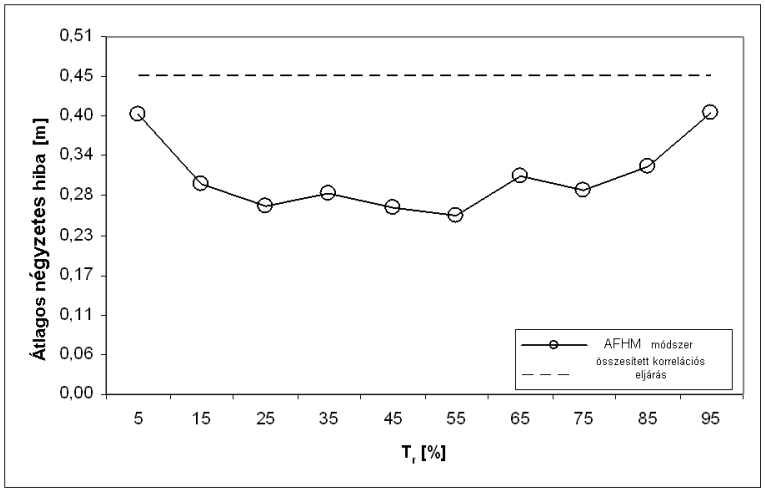
6.1. táblázat. A visszhanghatás-térképek elkészítéséhez szükséges idő Pentium IV. osztályú számítógépen.

95%-át meghaladó pontok alapján hoztam létre ($T_c = 0.95$). A módszer hatékonyságát a 6.8. ábrán látható hipotetikus előadó által bejárt utat feltételezve vizsgáltam. Az út első $(A_1 - A_2)$ szakaszán a beszélő a fal felé fordulva mozog az A_2 pont irányába, ezzel modellezve a táblára író és közben szüntelenül magyarázó előadó viselkedését. Az $(A_2 - A_3)$ valamint az $(A_3 - A_4)$ szakaszok megtétele közben az előadó a mozgásának megfelelő irányba beszél. A fenti út egyes pontjai kielégítik, míg mások sértik a 6.9. egyenletben leírt feltételt, így a kijelölt pontokkal vizsgálható a módszer hatékonysága azokban az esetekben, amikor a hagyományos forrás-lokalizáló algoritmusok hibás eredményre vezetnek a forrás iránykarakterisztikájának és a visszhangos környezet együttes hatása miatt, valamint azokban az esetekben is, amikor a 6.9. feltétel nem teljesül, tehát a hagyományos módszerek elméletileg helyes eredményt adhatnak.

6.5.2. A teljesítmény alakulása zajmentes esetben

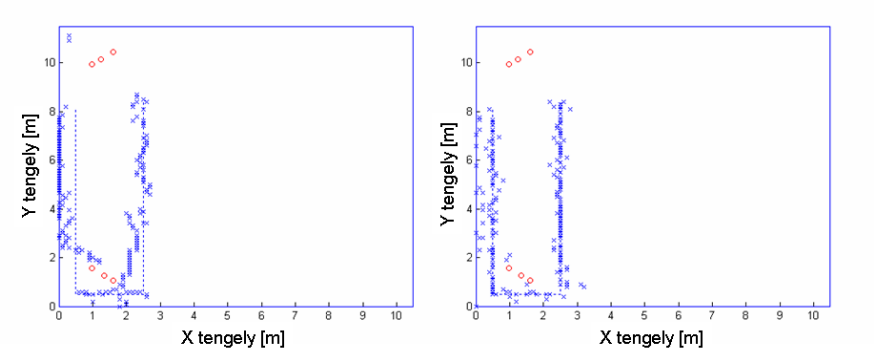
Annak érdekében, hogy az AFHM módszer teljesítményét ellenőrizzem, egy 27 másodperc hosszú, 25kHz-cel mintavételezett visszhangmentes felvételt készítettem a Budapesti Műszaki Egyetem Békésy György Akusztikai Laboratóriumában. A felvételt 40, egyenként 32768 mintát tartalmazó egymást körülbelül 50%-ban átfedő szegmensre osztottam. A mikrofonok szintetizált felvételeit, nyolcszoros visszaverődés figyelembevételével készített impulzusválasz-függvények konvolúciójával, ezen szegmenseket felhasználva állítottam elő, így modellezve a beszelő 6.8. ábrán feltüntetett mozgását. Az így elkészített felvételek 700Hz-es aluláteresztő szűréssel létrehozott változatainak segítségével összehasonlítottam az összesített korrelációs eljárás, valamint a bemutatott módszer hatékonyságát a vizsgált út 178 pontjában. Az AFHM módszer teljesítményét különböző figyelembe vett visszhanghatás értékeket (T_r) kiválasztva vizsgáltam meg. Az eredmények átlagos négyzetes hibája (ÁN hiba) a 6.9. ábrán látható.

Az eredményekből látható, hogy a bemutatott módszer hibája kisebb a összesített korrelációs eljárás hibájánál. A figyelembe vett visszhang optimális szintje a kísérletek szerint körülbelül 55%.



6.9. ábra. A vizsgált forrás-lokalizáló eljárások teljesítménye a 6.8. ábrán látható útvonalon.

E határ felett gyengébb teljesítményt kaptam, mivel a figyelembe vett visszhanghatások nem azonosítják egyértelműen a forrás helyét, azaz a becsült visszhanghatás-térképekről, valamint a megfigyelések alapján készített összesített korrelációs térképről olyan kevés lokális maximum helyet vettünk figyelembe a ponthalmazok távolságának meghatározásához, ami nem volt elegendő az egyes konfigurációk megkülönböztetéséhez. Amennyiben a gyengébb ($T_r = 15\%$ alatt) visszhanghatásokat is figyelembe vesszük, az AFHM módszer hatékonysága egyre csökken, mivel a korrelációs értékek természetes szórása miatt kialakuló csúcsok is visszhanghatásként értékeltek, vagyis már nem csak a visszhanghatások miatt létrejövő lokális maximumokat használjuk fel az akusztikus konfigurációk egyezésének vizsgálatára. A legoptimálisabb esetben is fennmaradó lokalizációs hiba több tényező együttes hatásának köszönhető. Ezek egyike, hogy az egymáshoz nagyon hasonló akusztikus konfigurációk visszhanghatás térképei közötti különbség eltünhet a térbeli diszkretizáció miatt. Másik probléma, hogy a visszhanghatásnak csak egy részét, a lokális maximumokat használjuk az akusztikus konfigurációk azonosítására, mi több ezen pontokból képzett halmazok súlypontja és inverz súlypontja alapján mérjük a megfigyelés és a becsült térképek közötti különbséget. A 6.10. ábrán látható részletes eredmények világosan mutatják, hogy jelentős teljesítmény különbség tapasztalható a két módszer között azokban az esetekben, amikor a 6.9. egyenletben leírt feltétel teljesül, míg a két módszer nagyjából azonos teljesítményt nyújt egyébként (részletesebben lásd a 6.2. táblázatban). Az AFHM módszer a 6.9. egyenletben leírt feltétel nem teljesülte esetén valamelyest gyengébb teljesítményt nyújt, aminek oka a ponthalmazok hasonlóság mérésének tökéletlensége.



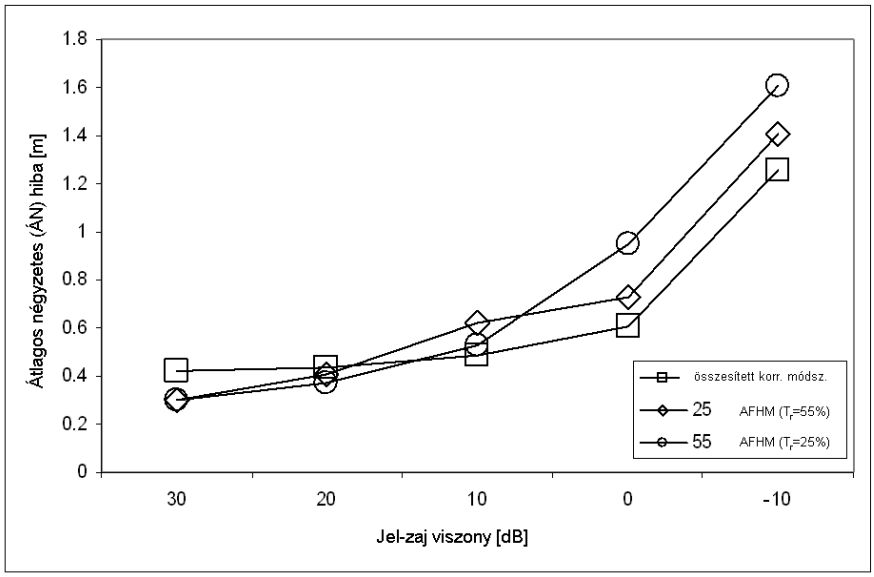
6.10. ábra. A forrás-lokalizáció eredményei. A bal oldali ábra az összesített korrelációs eljárás eredményét ábrázolja, míg az ábra jobb oldalán az AFHM algoritmus eredményei tekinthetőek meg $T_r = 0.55$ esetén.

	Átlagos négyzetes hiba [m] "valamely visszaverődési út direkt útnál kisebb csillapítása esetén"	Átlagos négyzetes hiba [m] "korrelációs maximum a direkt terjedési utak időkülönbségénél "
Pontok száma	134	44
Akkumulált/Összesített korrelációs eljárás	0.58	0
AFHM módszer (T_r=55%)	0.25	0.1
AFHM módszer (T_r=25%)	0.3	0.06

6.2. táblázat. A módszerek teljesítményének összehasonlítása a különböző terjedési utak függvényében.

6.5.3. A teljesítmény alakulása additív zajjal terhelt felvételek esetén

A forrás-lokalizáló módszerek zajjal szembeni robusztussága fontos szempont. Hasonlóan számos előzőleg publikált tanulmányhoz [127–129], jelen dolgozatban is élek azzal a feltételezéssel, hogy a mikrofonok által rögzített jelhez korrelálatlan fehér zaj adódik. A térben korreláló zaj ugyan jobban modellezi a valós életben előforduló zajforrások hatását, azonban a probléma bonyolultsága miatt ezidáig meglehetősen kevés munkában [130, 131] sikerült a létező módszerek lehetőségeit kiterjeszteni, éppen ezért jelen dolgozatban ezt a kérdést nem érintem. A kísérletben először előállítottam az előző fejezetben használt felvételek különböző jel-zaj viszonyú változatait a -10-től 30dB-ig terjedő intervallumban, majd ezek felhasználásával végeztem forrás-lokalizációt, mind az összesített korrelációs, mind az AFHM módszerrel.

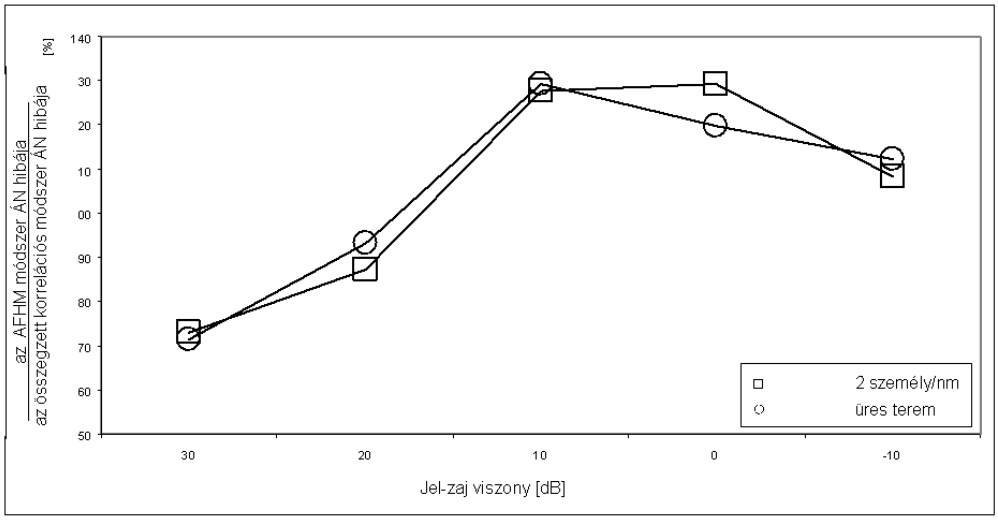


6.11. ábra. A fehér zaj hatása a lokalizáció teljesítményére.

A 6.11. ábrán közölt eredmények alapján elmondható, hogy az AFHM algoritmus, a figyelembe vett visszhanghatás-szintek mindegyikében ($T_r = 25\%$ és $T_r = 55\%$) érzékenyebb a jel-zaj viszony romlására mitöbb, már viszonylag magas jel-zaj viszony esetén is az összesített korrelációs módszer jobb teljesítményt nyújt a bemutatott eljárásnál. Ennek oka, hogy a visszhang hatásait lokális paraméterek formájában veszem figyelembe (egészen pontosan lokális maximumok formájában), ami az additív zaj okozta korrelációs tüskék miatt könnyen megbízhatatlan jellemzővé válik. Ennek a problémának egy lehetséges megoldása, hogy nem lokális paramétereket használunk a visszhang hatásainak követésére, hanem a visszhang által módosított tartományok (lásd 6.6. ábra) átlagát. Ennek elfogadható sebességgel történő számítása speciális hardver megoldásokat kíván. Ilyen lehet a Klefenz és kollégái [132] által bemutatott mesterséges Hubel-Wiesel hálózat, mellyel különböző görbületű vonalak detekciója valósítható meg valós időben.

6.5.4. Változó akusztikai körülmények vizsgálata

A forrás-lokalizáló módszerek hatékonyságának egyik kulcsa a visszhangos környezetben való alkalmazhatóság [86,95,96,98,99,101,104,127,128,133–136]. Mivel a tárgyalt módszer éppen a visszhang hatását használja fel a forrás helyének meghatározásához, a fenti kérdést jelen dolgozatban a változó akusztikai körülmények közötti viselkedésként értelmezem. Akusztikus környezetünket számos faktor [137] befolyásolja, úgy mint a levegő hőmérséklete, nedvességtartalma, vagy a visszaverő felületek elhelyezkedése és borítása. Konferenciatermi környezetet feltételezve a fenti faktorok jó közelítéssel állandónak tekinthetők, eltekintve a változó számú hallgatóságnak az akusztikus paraméterekre gyakorolt hatásától. Ennek vizsgálata érdekében az előző fejezetekben használt akusztikus modellt úgy módosítottam, hogy az eddig üresnek tekintett és a tömör fa visszaverődési tulajdonságaival modellezett széksorokat (6.7. ábra) a [138]-ben publikált adatoknak megfelelően négyzetméterenként két személy telítettségű nézőtér abszoprciós adataival helyettesítettem. Ennek következtében a terem utózengési ideje (T_{30}) az eddigi 3.5 másodpercről 1.5 másodpercre csökkent. A forráshely meghatározást a 6.5.1., illetve a 6.5.3. fejezetek szerint végeztem el, azzal a különbséggel, hogy a mikrofonjelek előállításához a hallgatókkal zsúfolt terem impulzusválasz-függvényét használtam fel. A kísérlet eredménye a 6.12. ábrán látható, a figyelembe vett visszhanghatás (T_r) 55%-os értéke esetén.



6.12. ábra. A módszer teljesítménye a becslések készítéséhez használt (*üres terem*) akusztikus modelltől eltérő (*2 személy/nm*) körülmények között.

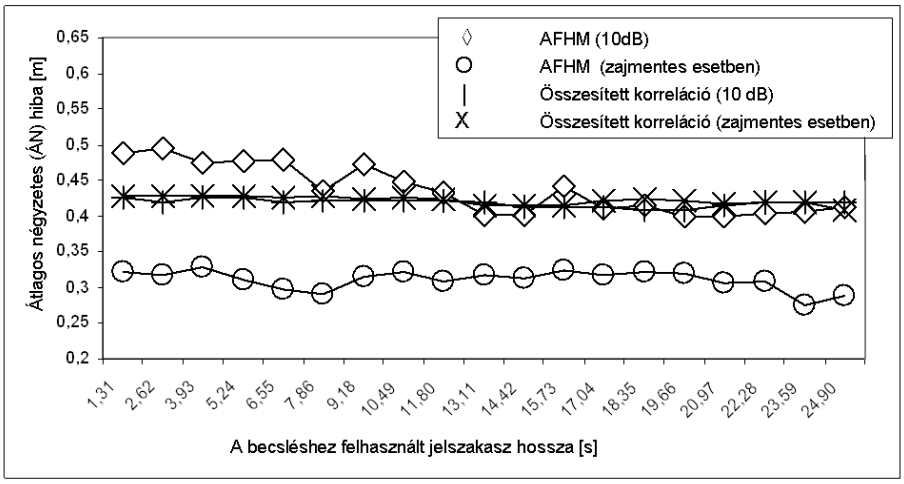
Látható, hogy a lokalizáció teljesítménye a terem telítettségének függvényében gyakorlatilag nem változott, a módszer tolerálja az akusztikus környezet mérsékelt változásából - a teljes visszaverő felület körülbelül 20%-át érintő jelentős abszorpciós képesség változásból - fakadó hatásokat.

6.5.5. Az módszer konvergenciája

Mivel a hangforrások helyének meghatározásával foglalkozó munkákban a rossz minőségű felvételek alapján történő lokalizációt hagyományosan az egyes becslések eredményének alkalmas összegzésével próbálják javítani, érdemes megvizsgálni az eredmények valódi forráshelyhez való konvergenciájának sebességét. A szóban forgó eljárás esetén a becslések aggregációját az összesített korrelációs térképek akkumulációján keresztül valósítom meg, ezért új jelöléseket vezetek be a mérések pillanatnyi eredménye alapján készített összesített korrelációs térképre:

$$£(l) = \sum_{i=L-S}^{L} £_i(l) \qquad \forall l \in C \tag{6.20}$$

ahol $£_i(l)$ jelöli az i. mérés alapján, az 5.15. egyenlet szerint számított összesített korrelációs térkép l pontban felvett értékét, L a legutolsó mérést jelöli, S pedig az előző mérések alapján készített korrelációs térképek számáa. S értékét a felhasználás sajátosságainak megfelelően kell megválasztani, például a hangforrás maximális sebességétől, a mintavételi időtől, a korrelációs ablak méretétől (W) függően. A dolgozatban közölt kísérlet során az $S = L$ értékkel számoltam, mivel nem kívántam alulról korlátozni a konvergencia sebességét. Az eddig használt akusztikus modellt felhasználva a 6.8. ábrán látható elrendezésen ellenőriztük a konvergencia sebsséget, azonban az előzőekkel ellentétben ezúttal a beszélő a jelzett út minden pontjában eltöltött 27 másodpercet, így a forrás helyére vonatkozóan pontonként negyven becslést olvashattunk ki. Az egyes időpillanatokhoz tartozó négyzetes hibák átlaga alapján kaptuk a 6.13. ábrán látható diagramot.



6.13. ábra. A módszerek konvergenciasebesség vizsgálatának eredménye.

Zajmentes esetben az eredmények a vártnak megfelelően azt igazolják, hogy az összesített korrelációs módszer teljesítményét a mérések eredményének időbeni átlaga nem befolyásolja, hiszen a lokalizációs hiba a forrás anizotrop karakterisztikájából fakad, erre pedig a mérési eredmények átlagolása nincs hatással. A zaj által okozott hiba az összesített korrelációs módszer esetén a vizsgált jel-zaj viszony érték mellett elenyésző, ezért ez a görbe is közel állandó hibát jelez. Érdekesebb

következtetés vonható le a tárgyalt módszer eredményeinek vizsgálatával, mivel az egyes becslések eredményeinek átlagát felhasználva a periodicitásból fakadó nem kívánt korrelációs csúcsok hatása csökkenthető. A zajmentes jellel kapott eredmények bizonyítják, hogy a figyelembe vett jelszakasz aperiódikus volta nem javítja tovább az algoritmus teljesítményét, tehát a PHAT súlyozás alkalmazása indokolt. Ez alapján állíthatjuk, hogy a fennmaradó lokalizációs hiba a becslések és a megfigyelések közötti hasonlóságmérték hibájából, illetve a térbeli diszkretizáció okozta bizonytalanságból fakad. A 10dB jel-zaj viszonyú mérések felhasználásával kapott eredmények igazolják, hogy a zaj által okozott lokális maximumok a becslések átlagát véve kioltják egymást, és csak azok a maximumok maradnak meg, amelyek a visszhanghatásnak tulajdoníthatóak. Az adatsorok alaposabb vizsgálata azonban azt bizonyítja, hogy a konvergencia lassú, a 40 szegmens feldolgozását követően is csak az összesített korrelációs módszer teljesítményével vethető össze.

6.6. Diszkusszió

6.6.1. Az alkalmazott akusztikus modell érvényessége

Az AFHM módszer frekvenciafüggetlen geometriai hangterjedési modellje csak bizonyos feltételek teljesülte esetén tekinthető a valós körülmények jó közelítésének. Ezen megkötések az alábbiak:

- A hang hullámhossza lényegesen kisebb, mint a visszaverő felületek kiterjedése.
- A visszaverő felületek síknak tekinthetőek a hang hullámhosszához viszonyítva.
- A modellbe integrált visszaverő felületeket kivéve a hang terjedésének útjában nincs a hang hullámhosszával összemérhető kiterjedésű objektum.

Azokban az esetekben amikor az első és a harmadik feltétel teljesül a hang hullámok diffrakciója következik be, míg a második feltétel sérülése az úgynevezett szóródás[5] hatással modellezhető. Tipikus konferenciatermi alkalmazásokat tekintve a harmadik kritérium teljesülése kellően konzervatív feltételezés. Ugyanez nem mondható el a fennmaradó faktorok hatásáról, melyek kielégítő hatékonyságú számítógépes modelljeinek elkészítése aktív kutatási terület. A vonatkozó munkák [94,139,140] azt sugallják, hogy a legkorábbi visszaverődések jól modellezhetők geometriai hangterjedéssel. Mivel a legelső visszaverődések tartalmazzák az energia nagy részét, a módszer alkalmas a legnagyobb kereszt-korrelációs csúcsok predikciójára. Az alkalmazási környezet tipikus méretei alapján megjósolható, hogy az alkalmazott akusztikus modell mely frekvenciákon ad jó becslést a valóságos terjedési és visszaverődési jelenségekre. A 6.3. táblázatban, négy tipikus teremméret esetére határoztam meg azt az alsó frekvenciát, ameddig a geometriai akusztikus modellek jó közelítésnek tekinthetőek. A táblázatban közölt eredmények mutatják, hogy a viszonylag kis térfogatú zárt terekben csak a beszéd magas frekvencia-komponensei használhatóak, következésképpen az AFHM módszer jellemző felhasználási területe az előadó-, illetve konferenciatermekben való forrás-lokalizáció.

6.6.2. A módszer számításigénye

A forrás-lokalizáló algoritmusok sebessége elsődleges fontosságú, mivel jellemzően valós időben van szükség a forrás helyének azonosítására. A 6.4. táblázatban az összesített korrelációs eljárás [86], az illeszkedő szűrőtömbökkel kiegészített nyalábirányítási technika [114], valamint az ASL módszer előzetes és valós időben számítandó feladatait összegeztük.

[5] scattering

Jellemző felhasználási környezet	Tipikus kiterjedés (magasság · szélesség · mélység)	Alsó frekvencia határ
Iroda	3m · 5m · 5m	2 kHz
Tanterem	3m · 10m · 6m	1.5 kHz
Kis előadó	5m · 15m · 10m	600 Hz
Konferencia terem	8m · 30m · 30m	200 Hz

6.3. táblázat. A geometriai akusztikus modellek használhatóságának alsó frekvencia határa tipikus teremméretek esetén [141].

Az ASL módszer vitathatatlan előnye az illesztett szűrőtömbökön alapuló módszerhez képest, hogy nem szükséges a jelek megfelelő helyre vonatkozó impulzusválasz-függvénnyel történő dekonvolúciója, mivel a visszhanghatások előzetesen, a visszhanghatás-térképet elkészítésekor kerülnek kiértékelésre. Másrészt az ASL eljárás összesített korrelációs módszerhez viszonyított többlet számításigénye egy lokális maximum keresés és a lehetséges konfigurációk halmazába (fC) tartozó ponthalmazok hasonlóságának meghatározása. A kísérletek során a lehetséges konfigurációk halmazába tartozó akusztikus konfigurációk száma egyetlen esetben sem haladta meg 100-at, ezért a gradiens keresés teszi ki számítási igény többlet nagy részét, ami nem számottevő különbség.

Algoritmus	Előzetesen elvégzendő feladatok	Valós idejű számításigény
összesített korrelációs módszer	-	- a kereszt-korrelációs függvények meghatározása - közös koordináta-rendszerbe vetítés
nyalábirányítás illesztett szűrőtömbökkel	- az impulzusválasz-függvény meghatározása minden lehetséges forráshelyre	- a beérkező jel megfelelő helyhez tartozó impulzusválasz-függvény szerinti dekonvolúciója minden lehetséges forráshelyre - a dekonvolúcióval képzett jelen a nyalábirányítás eredményének meghatározása
ASL módszer	- az impulzusválasz-függvény meghatározása minden lehetséges akusztikus konfigurációra - a becsült visszhanghatás-térképek elkészítése minden lehetséges akusztikus konfigurációra - A lokális maximumok megkeresése minden becsült visszhanghatás-térkép esetén	- a kereszt-korrelációs függvények meghatározása - közös koordináta-rendszerbe vetítés - lokális maximum keresés - a tárolt konfigurációk és a megfigyelés alapján létrehozott térkép hasonlóságának mérése

6.4. táblázat. Az egyes algoritmusok esetén valós időben, illetve előzetesen számítandó feladatok.

7. FEJEZET

Konkluzió és a további feladatok

7.1. Áttekintés

A dolgozatban konkurens akusztikus források jeleinek szétválasztására használható algoritmusokat mutattam be.

A disszertáció első részében a hangok fizikai jellemzők szerinti, heurisztikus módszerekkel történő szétválasztását tárgyaltam, mely módszerek mintájául az emberi hallórendszer pszichoakusztikus megfigyelésekkel azonosított csoportosítási szabályai szolgálnak. Az eljárásokat egy celluláris hullámszámítógépen alkalmazható programkönyvtár részeként ismertettem. A felhasználás módjára egy példa alkalmazást mutattam, amelyben azonos időben beszélő személyek, hang alapján történő helymeghatározásának hibáját sikerült jelentősen csökkenteni.

A forrás-szeparációs probléma megoldásának egy másik stratégiája a források különböző térbeli elhelyezkedése alapján megvalósított szegregáció. Áttekintettem a forrás-lokalizációs feladatok megoldásában alkalmazható algoritmusokat, majd rámutattam, hogy visszhangos környezetben a forrás anizotrop tulajdonságából fakadóan a hagyományos érkezési-időkülönbség becslő algoritmusok hibás eredményre vezetnek. Bemutattam egy, az akusztikus környezet hatásait figyelembe vevő forrás-lokalizáló eljárást, mely zajmentes esetben a közölt algoritmusoknál lényegesen hatékonyabban képes a forrás helyének meghatározására. Vizsgáltam a közölt algoritmus változó akusztikus körülmények között való felhasználásának lehetőségeit, illetve a számítási igényt figyelembe véve összehasonlítást végeztem más korszerű forrás-lokalizáló módszerekkel.

7.2. Módszerek, eszközök

A dolgozatban közölt módszerek interdiszciplináris kutatómunka eredményei, melyek koherensen ötvözik a teremakusztika, a pszichoakusztika, a Celluláris Neurális Hálózatok, valamint a jelfeldolgozás területéről származó ismereteket.
Kutatásaim során a konkurens források időbeni szegregációjával kapcsolatos kísérletek végrehajtása érdekében létrehoztam egy hatékonyan számítható és a kísérletek szempontjából releváns információkat megőrző, valamint azokat kiemelő, a cochlea funkcionális modellje alapján készített cochlea szimulátort. A szimulátorral előállított kétdimenziós spektro-temporális képfolyamon alkalmaztam a *hallási jelenet elemzés* elméletéből ismert csoportosítási algoritmusok Celluláris Hullámszámítógépen

futó megvalósításait. A Celluláris Hullámszámítógépen történő implementáció során a feladat megoldását célzó algoritmusok létrehozásakor különös gonddal vettem figyelembe a létező CNN-UM implementációk támasztotta követelményeket. A felhasznált template-ek kiválasztásánál a CNN Software Library-t használtam referenciaként, ügyelve arra, hogy a kiválasztott template, hardver környezetben való felhasználására létező és robusztus megoldások álljanak rendelkezésre. Azokban az esetekben, ahol a kívánt feladat megoldását célzó súlymátrixok nem álltak rendelkezésre, a parciális differenciálegyenletekre vonatkozó tételeket és állításokat felhasználva hoztam létre új template-eket, ellenőrizve a stabilitásra, a robusztusságra és a különböző CNN-UM platformokon történő megvalósíthatóságra vonatkozó szempontokat. A pszichoakusztikus modellkönyvtárat az AladdinPro szoftver szimulátort használva fejlesztettem ki. Az elkészült AMC forrás file-okat szabadon felhasználható mintaként, az algoritmusok dokumentációját UMF leírásban tettem hozzáférhetővé. A különböző platformok közötti átjárhatóságot biztosító segédprogramokat Matlab-ban készítettem el.
A hangforrások térbeli szegregációjának és elhelyezkedésének vizsgálatához a hang geometriai terjedésén alapuló modellt használtam. Tanulmányoztam a modell érvényességének határait, majd a matematikai analízis és a jelfeldolgozás eszközeit felhasználva következtetéseket fogalmaztam meg visszhangos környezetben elhelyezett anizotrop források hagyományos forráslokalizáló algoritmusokra gyakorolt hatására. A valószínűség-számítás eszközeit felhasználva becsülhetővé tettem a forrás helyére jellemző, a visszhang hatásaként létrejövő kereszt-korrelációs csúcsokat, majd a gépi-tanulás területéről származó tapasztalatokat felhasználva módszert adtam a megfigyelésekhez legjobban illeszkedő konfiguráció kiválasztására. A kidolgozott módszert teljesítményét C++-ban implementált rutinok segítségével a CAT akusztikus modellező szoftvert felhasználva ellenőriztem.

7.3. Tudományos eredmények

1. Tézis csoport
Kialakítottam egy hullámszámítási keretrendszert, mely az emberi hallórendszer néhány aspektusát hatékonyan modellezi. A keretrendszer a cochlea funkcionális analógiáján alapuló frekvencia-felbontással előállított kétdimenziós spektro-temporális folyamnak a *hallási jelenet elemzés* elméletéből ismert sajátosságok szerinti feldolgozásához szükséges analogikai algoritmusokat tartalmazza.

1.1. A természetben előforduló fizikai folyamatok által keltett hangjelenségek sajátossága, hogy spektrális komponenseik minden tagjában azonos időben jelenik meg a kisugárzott energia. Új hullámszámítási algoritmust dolgoztam ki a „szinkron kezdet" csoportosítási szabály mintájáramegvalósítására. A kidolgozott algoritmus a kétdimenziós frekvencia-idő hangképen bináris hullámok ütközése révén, logikai műveletek segítségével azonosítja a különböző frekvenciatartományokban azonos időben megjelenő komponenseket.

1.2. A természetes folyamatok által keltett hangok spektrális tartalma általában azonos módon változik. Az azonos módon változó - közös frekvencia és/vagy amplitúdó modulált - hangjeleket hallórendszerünk azonos forrásból érkező hang objektumként azonosítja. Módszereket adtam azonos sorsú, azaz közös amplitudó-, illetve frekvencia-modulációjú jelek analogikai algoritmussal történő azonosítására.
A közös amplitúdó modulált jelek kiválasztását időben szinkron kezdetű és végű jelek kiválasztásának problémájára vezettem vissza, felhasználva az előző tézispont eredményeit.

A közös frekvencia moduláció hatása a cochleáris transzformáció sajátosságából fakadóan az egyes frekvencia-sávok energiatartalmának állandó spektrális távolságaként jelenik meg. A kidolgozott analogikai algoritmus az állandó spektrális távolság meglétét egy új, robusztus NxN-es template osztály alkalmazásával ellenőrzi, mely lineáris lépésben dekomponálható 3x3-as template-szekvenciává, lehetővé téve a szilícium alapú CNN-UM implementációkon való alkalmazást.

1.3. A hangforrások a kisugárzott hangenergiát rövid megszakítást követően, egy az addigi frekvenciához közeli sávban sugározhatják tovább. A cochleáris modell kimenetén a fenti jelenség rövid „réseket" eredményez. A bináris hullámok számítási lehetőségeit kiaknázva kidolgoztam a „folytonosság" pszichoakusztikus csoportosítási szabálynak megfelelő eljárást, mely lineáris időben jelöli ki a meghatározott paramétereknek eleget tevő területeket, így hozva létre egységes hangobjektumokat.

1.4. Hallórendszerünk az egymáshoz frekvenciában és időben közeli energia komponenseket közös hangobjektumként kezeli. Eljárást dolgoztam ki, mely az alkalmazott celluláris struktúrának köszönhetően hatékonyan emeli ki a meghatározott energiaátlag feletti területeket, így alakítva ki a spektrális és időbeni távolság alapján szerveződő „közelség" csoportosítási szabállyal azonosított hangobjektumokat.

Kapcsolódó közlemény:

***Z. Fodróczi**, A. Radványi „Computational Auditory Scene Analysis in Cellular Wave Computing Framework" International Journal of Circuit Theory and Applications Vol: 34(4) pp: 489-515, ISSN:0098-9886 (July 2006)*

2. Tézis csoport
Új forrás-lokalizáló eljárást dolgoztam ki, amivel zajmentes körülmények közt a hagyományos algoritmusoknál lényegesen hatékonyabban határozható meg visszhangos környezetbe helyezett anizotrop források helye. A módszer a geometriai hangterjedés-modell segítségével az akusztikus környezet és a forrás iránykarakterisztika együttes hatását figyelembevéve határozza meg a hangforrás helyét. Az eljárással speciális cél-hardver nélkül, az előzetesen végrehajtott akusztikus számítások eredményeit felhasználva valós időben végezhető forrás-lokalizáció.

2.1. Az alkalmazott akusztikus modell segítségével megadtam a visszhangos környezetben elhelyezett pontszerű forrás hangját rögzítő mikrofonok jeleinek időfüggvényét. Ezeket felhasználva auto-korrelációs függvények lineáris kombinációjaként felírtam tetszőleges mikrofonpár kereszt-korrelációs függvényét. Az auto-korrelációs függvény tulajdonságait megvizsgálva becslést adtam az akusztikus környezet által a kereszt-korrelációs függvényre gyakorolt hatásra.

2.2. A kidolgozott modell keretei között vizsgáltam a visszhangos környezetbe helyezett anizotrop forrás kereszt-korrelációs függvényre gyakorolt hatását. Feltételt fogalmaztam meg, melynek sérülése esetén a forrás iránykarakterisztika és az akusztikus környezet együttes hatása miatt, a hagyományos érkezési-időkülönbség becslő eljárások a forráshely meghatározására alkalmatlanná válnak.

2.3. Az *összegzett korrelációs térkép* eljárás adaptációjával becsült visszhanghatás-térképeket hoztam létre, melyekkel a mikrofonpáronként becsült visszhanghatás hatékony és robusztus

összegzését valósítottam meg. A becsült visszhanghatás-térképek lokális maximum helyeinek meghatározásával, az alkalmazott akusztikus konfigurációt jellemző négydimenziós ponthalmazokat hoztam létre.

2.4. Eljárást adtam a megfigyelés alapján készített összegzett korrelációs térkép visszhanghatásainak kinyerésére, majd az így nyert ponthalmazt felhasználva távolság mértéket definiáltam a megfigyelések és a becsült visszhanghatás-térképek hasonlóságának kifejezésére. A létrehozott hasonlóság mérték segítségével zajmentes körülmények között azonosítható, hogy a tárolt konfigurációk közül melyik a megfigyelésekhez legjobban illeszkedő, így adva becslést a forrás hipotetikus helyére.

Kapcsolódó közlemény:

***Z. Fodróczi**, A Radványi. „Localization of Directional Sound Sources Supported by a priori Information of the Acoustic Environment" manuscript accepted to EURASIP Journal on Applied Signal Processing*

7.4. Az eredmények alkalmazási területei

A tézisekben bemutatott algoritmusok konkurens források jeleinek szétválasztására használhatóak. A forrásonként szegregált jelek az első tézisben bemutatott megoldással közvetlenül felhasználhatóak a megfelelő jelszegmensek előzetes kiválasztása révén a forrás-lokalizáló algoritmusok hibájának csökkentésére. A szegregált jelek további felhasználási területe a mesterséges beszéd, illetve hangesemény felismerő rendszerek teljesítményének növelése, mivel a jelenleg ismert algoritmusok rendkívül érzékenyek a felismerési feladathoz nem kapcsolódó egyéb nem kívánatos hanghatások jelenlétére. A bemutatott módszerrel a valós életben előforduló koktélparti effektusként említett helyzetek jó része term;szetesen nem oldható meg az emberi hallgatók teljesítményéhez fogható eredménnyel. Az elmúlt több mint 30 évben megoldhatatlannak talált feladatra tehát ezúttal sem sikerült minden szempontból kielégítő választ találni. A munkám eredménye azonban az, hogy rámutat, hogy az alternatív számítási paradigmák alkalmazásával elérhető nagy számítási teljesítmény közelebb visz a megoldáshoz azáltal, hogy a tanult, illetve sémavezérelt, magasabb hierarchiájú folyamatok által irányított a primitív csoportosítási szabályok adaptív-iteratív kiértékelése valós időben válik lehetségessé.

A teremalkalmazásokon túlmutató lehetőséget rejt - a feladathoz alkalmasan megválasztott architektúra esetén - a nagy számítási teljesítmény mellett elérhető alacsony energiafogyasztás, aminek révén a kidolgozott algoritmusokkal hallókészülékek, illetve cochlea protézisek adaptív és kontextus függő vezérlése valósítható meg. A második tézisben bemutatott algoritmus segítségével beszélők helyének biztosabb meghatározása válik lehetségessé, ami közvetlenül hathat biztonsági megfigyelő hálózatok és automatikus videokonferencia rendszerek hatékonyságára. Emellett a beszélők helyének pontosabb meghatározása irányított mikrofontömbök alkalmazása révén tisztább, a beszélő hangját jobban kiemelő felvételek készítését biztosítja, ami a mesterséges beszédfelismerő rendszerek teljesítményének növekedését eredményezi. A bemutatott módszerekkel a valós életben előforduló koktélparti effektusként említett helyzetek jó része valószínűleg nem oldható meg az emberi hallgatók teljesítményéhez fogható eredménnyel. Az elmúlt több mint 30 évben megoldhatatlannak talált feladatra tehát nekem sem sikerült minden szempontból kielégítő választ találni. A dolgozat szándékolt célja az, hogy rámutasson, hogy az alternatív számítási paradigmák alkalmazásával

elérhető nagy számítási teljesítmény közelebb visz a megoldáshoz azáltal, hogy a tanult, illetve sémavezérelt, magasabb hierarchiájú folyamatok által irányított a primitív csoportosítási szabályok adaptív-iteratív kiértékelése valós időben válik lehetségessé. Amennyiben a jel feldolgozása megfelel?en rövid id? alatt kivitelezhet?, a mozgó források nem jelentenek problémát.

7.5. A további kutatás lehetséges irányai

A forrás-szeparációs probléma megoldása hosszú évtizedek óta kutatott terület. A napjainkban is meglevő nehézségek megoldására tett újabb és újabb erőfeszítéseket a biológiai rendszerek zavarbaejtő képességeinek „egzisztencia bizonyítéka" tartja életben. Az elmúlt több mint 30 év eredményei azonban számvetésre késztetnek, hiszen ma sem rendelkezünk az előlények képességeit akár csak megközelítő műszaki megoldásokkal. A forrás-lokalizáció problémáját megoldani hivatott algoritmusok a jelfeldolgozás igen összetett és figyelemreméltó elméleti eredményei ellenére sem képesek értékelhető választ adni a mindennapi életben tapasztalható zaj, visszhang és egyéb hatások jelenlétében. Léteznek a biológiai rendszerek forrás-lokalizációval kapcsolatba hozható idegi struktúráinak analógiája alapján működő megoldások is, ezek azonban nem jelentenek minőségi változást, mivel a megoldás filozófiáját tekintve ugyanazt az elgondolást követik, mint a jelfeldolgozás eszközeit alkalmazó megoldások.

7.5.1. A forrás-lokalizációs probléma

Mint arra a 6. fejezetben rámutatok, a forrás-lokalizációs probléma pusztán a szenzorokhoz érkező jelek időkülönbségének azonosításával nem oldható meg, hiszen a forrás anizotrop tulajdonsága és a visszhang együttes hatása szükségszerűen vezethet hibás helymeghatározáshoz. Elengedhetetlen tehát akár a környezet akusztikus hatásait figyelembe vevő, akár azok hatását kiszűrni képes megoldások kidolgozása.

A dolgozat 6. fejezetében e hatások integrációjára mutattam példát. A módszer meglevő hibáit kiküszöbölendő a jövőben érdemes lenne megvizsgálni a visszhanghatások globális paraméterek alapján való figyelembevételének módját, amihez z kapcsolódóan a 6.5.3. fejezetben olvashatóak gondolatok. Nagyban szélesítené az algoritmus alkalmazási lehetőségeit a visszhanghatás becslések több frekvenciatartományra való elkészítése, ami lehetővé tenné a rögzített jel spektrális tartalmához jobban illeszkedő becslések kiválasztását.

Az akusztikus környezet impulzusválasz-függvényeinek explicit meghatározásán alapuló megoldások figyelemre méltó alternatívái a függvények iteratív becslésével kísérletező eljárások, melyek az 5.4.3. fejezetben tárgyalt módszerek közé sorolhatóak. A módszerek egyelőre zajérzékenyek, illetve nem tisztázott a több mikrofonpárt érintő adaptív optimalizációs probléma megoldásának módja sem.

A szigorúan vett jelfeldolgozásnál valamivel messzebb vezet annak vizsgálata, hogy az élőlények testtartásának, illetve fejállásának akusztikus teret befolyásoló hatása mekkora szerepet játszik a forrás helyének meghatározásában. Valószínűsíthető, hogy az előlények megtanulják, hogy a különböző irányból érkező hangok spektrális tartalma különböző fejállás esetén milyen változáson megy keresztül. Ez a jellemző fontos kiegészítője lehet az érkezési-időkülönbség becslő algoritmusoknak.

7.5.2. Kontextuális információval segített forrás szeparáció

A bemutatott *hallási jelenet elemzés* könyvtár egyik fontos továbbfejleszési lehetősége a zajjal szembeni érzékenység vizsgálata, illetve annak növelése. Ennek egyik módja lehet a kidolgozott cochleáris transzformáció adaptívitásának továbbfejlesztése, valamint az egyes szabályok implementációját érintő, a 4.2.6. fejezetben megfogalmazott gondolatok. Mint arra a 2. fejezetben utaltam a szegregációjában nagy valószínűséggel fontos szerepet játszanak a kibocsátott hangok egyes tulajdonságaira vonatkozó előzetes ismeretek, melyek a sémavezérelt csoportosítási mechanizmusokon keresztül fejtik ki hatásukat. Ilyen lehet a kibocsátó forrás ismert viselkedéséből származó információ, például egy elhaladó gépkocsi hangjának egyéb forrásoktól való elkülönítése esetén. A legjelentősebb azonban a már azonosított forrásoktól függő kontextusban végzett asszociatív felismerés. E funkciónak köszönhető, hogy képesek vagyunk nagy háttérzajban is kiválasztani a minket érdeklő forrásból érkező információt. A felismert kontextusnak köszönhetően, a zajos, gyakran sérült vagy deformált jeleket csak néhány hipotézis ellenőrzésére kell felhasználnunk. Egyelőre nem világos, hogy a sémavezérelt mechanizmusok milyen módon befolyásolják az adatvezérelt csoportosítási szabályok kiértékelését. Valószínű, hogy az adatvezérelt csoportosítási szabályok kiértékelése már ugyancsak egy valamelyest szűkített kontextus értelmezésének fényében, viszonylag egyszerű, alacsony szintű, prediktív modellekkel segítve történik. A kognitív idegtudomány egyik figyelemre méltó hipotézise, hogy ezen prediktív modellek aktualizálása EEG elektródákkal mérhető változást, az eseményhez kötött potenciál[1] kiváltását okozza. E jelenség természetére vonatkozóan viszonylag sok információ áll rendelkezésre, illetve további kísérletekkel információt szerezhetünk a prediktív modellek működéséről, ezért időszerű egy analóg számítógépes modell építése, mely nélkülözhetetlen része lehet a jövő hangfeldolgozó rendszereinek.

[1] event related potencial

A szerző publikációi

Folyóirat publikációk:

Z. Fodróczi, A. Radványi „Computational Auditory Scene Analysis in Cellular Wave Computing Framework" International Journal of Circuit Theory and Applications Vol: 34(4) pp: 489-515, ISSN:0098-9886 (July 2006)

Z. Fodróczi, A Radványi. „Localization of Directional Sound Sources Supported by a priori Information of the Acoustic Environment" manuscript accepted to EURASIP Journal on Applied Signal Processing

Konferencia előadások:

Z. Fodróczi, A. Radványi, Gy. Takács „Acoustic Source Localization using Microphone Arrays via CNN algorithms" Proceedings of 3rd International Conference on European Conference on Circuit Theory and Design (ECCTD03) 2003

Könyv fejezetek:

Á. Novák, A. Sali, K. Kis, **Z. Fodróczi** „First Course On Database Management System" - „Structured Query Language" Chapter 5; „Be wired - Intoduction into HTML and PHP" Chapter 16; „eXtended Markup Language" Chapter 17 edited by Á. Novák

Irodalomjegyzék

[1] Mark D. Skowronski and John G. Harris. Human factor cepstral coefficients: Biological inspiration + engineering = noise-robust speech features. In *in Proceedings of the Acoustical Society of America First Pan-American/Iberian Meeting on Acoustics*, 2002.

[2] E.C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *Journal of Acoustic Society of America*, 25:975–979, 1953.

[3] I. Winkler. *Modell-vezérelt folyamatok a hallási környezet leképezésében.* PhD thesis, MTA Pszichológiai Intézet, Budapest, Hugary, 2004.

[4] Max Wertheimer, Wolfgang Köhler, and Kurt Koffka. Experimentelle studien über das sehen von bewegung. *Zeitschrift für Psychologie*, 61:161–265, 1922.

[5] S. Lehar. *The World In Your Head.* Lawrence Erlbaum, Mahwah, NJ., 2003.

[6] Albert S. Bregman. *Auditory Scene Analysis.* MIT Press, Cambridge, 1990.

[7] M. Cooke. *Modelling Auditory Processing and Organization.* PhD thesis, The University of Sheffeld, Sheffield, England, 1991.

[8] M. Cooke and G. Brown. Computational auditory scene analysis: Exploiting principles of perceived continuity. *Speech Communication*, 13:391–399, 1993.

[9] D. P. W Ellis. A computer implementation of psychoacoustic grouping rules. *in Proceedings of the 12th International Conference on Pattern Recognition*, 1994.

[10] D. Ellis. *Prediction driven computational auditory scene analysis.* PhD thesis, MIT, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 1997.

[11] R. M. Warren. Restoration of missing speech sounds. *Science*, 1970.

[12] Wersenyi Gy. *HRTFs in Human Localization: Measurement, Spectral Evaluation and Practical Use in Virtaul Audio Environment.* PhD thesis, Brandenburgische Technische Universitaet, NJ, Cottbus, Germany, 2002.

[13] A. Fonyó. *Az orvosi élettan tankönyve.* Medicina Könyvkiadó Rt., Budapest, 1999.

[14] Gy. Bekesy. *Experiments in Hearing.* McGraw Hill Book Co., New York, 1960.

[15] D.L. Oliver, G.E. Beckius, D.C. Bishop, W.C. Loftus, and R. Batra. Topography of Interaural Temporal Disparity Coding in Projections of Medial Superior Olive to Inferior Colliculus. *Journal of Neuroscience*, 23(19):7438–7449, 2003.

[16] G. Brown and M. Cooke. Computational auditory scene analysis. *Computer Speech and Language*, 8:297–336, 1994.

[17] G. J. Brown and M. Cooke. Perceptual grouping of musical sounds: A computational model. *The Journal of New Music Research*, 23:107–132, 1994.

[18] G. Brown and M. Cooke. Temporal syncronization in neural oscillatory model of primitiv auditory stream segregation. *In working notes of the Workshop on Computational Auditory Scene Analysis at the International Conference of Artificial Intelligence*, pages 41–47, 1995.

[19] G. Brown. *Computational Auditory Scene Analysis: A Representational Approach.* PhD thesis, The University of Sheffeld, Sheffeld, England, 1992.

[20] G. J. Brown and D. Wang. *Speech enhancement.* Springer, New York, 2005.

[21] P. Denibgh and J. Zhao. Pitch extraction and separation of overlapping speech. *Speech Communication*, 11:119–125, 1992.

[22] D. J. Godsmark and G.J. Brown. Context-sensitive selection of competing auditory organisations: a blackboard model. In *In working notes of the Workshop on Computational Auditory Scene Analysis at the Internationa Joint Conference on Artificial Intelligence*, pages 60–67, Montreal, 1995.

[23] [14] Guoning H and DeLiang W. Auditory segmentation based on onset and offset analysis. *Technical Report OSU-GSRC-1/05-TR04*, 2005.

[24] DeLiang W Guoning H. Separation of stop consonants. *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP.03)*, 2003.

[25] L. A. Drake. *Sound source separation via computational auditory scene analysis (casa)-enhanced beamforming.* PhD thesis, Northwestern University, Evanston, Ilinois, 2001.

[26] D. Mellingger. *Event formation and separation in musical sounds.* PhD thesis, Standford University, 1991.

[27] T. Nakatani and H. Okuno. Harmonic sound stream segregation using localization and its application to speech stream segregation. *Speech Communication*, 27:209–222, 1999.

[28] H.G. Kuno, T. Nakatani, and T. Kawabata. Auditory stream segregation in auditory scene analysis with multi-agent system. *in Proceedings of American Association of Artificial Intelligence*, 1994.

[29] T. Nakatani, H. Okuno, and T. Kawabata. Residue-driven architecture for computational auditory scene analysis. *in Proceedings of 14th International Joint Conference on Artificial Intelligence, (IJCAI-95)*, pages 165–172, 1995.

[30] Paris Smaragdis. *Redundancy Reduction for Computational Audition, a Unifying Approach.* PhD thesis, MIT, Massachusetts Institute of Technology, 2001.

[31] T. W. Parison. Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America*, 60:911–918, 1976.

[32] M. Karjalainen and T. Tolonen. Multi-pitch and periodicity analysis model for sound separation andauditory scene analysis. *in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'99)*, 2, 1999.

[33] D. F. Rosenthal and H. G. Okuno. *Computational auditory scene analysis.* London, Erlbaum, 1997.

[34] G. Hu and D.L. Wang. Auditory segmentation based on onset and offset analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, page in press, 2006.

[35] D. L. Wang and G. J. Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transaction on Neural Networks*, 10:684–697, 1999.

[36] N. Roman and D. L. Wang. Pitch-based monaural segregation of reverberant speech. *Journal of the Acoustical Society of America*, page in press, 2006.

[37] M. Weintraub. *A theory and computational model of auditory sound separation.* PhD thesis, Standford University, 1985.

[38] S. N. Wrigley. *A theory and Computational Model of Auditory Selective Attention.* PhD thesis, The University of Sheffeld, Sheffeld, England, 2002.

[39] H. J. Nussbaumer. Fast Fourier transform and convolution algorithms. *Berlin and New York, Springer-Verlag(Springer Series in Information Sciences.*, 2, 1982.

[40] R. Patterson and B. Moore. *Auditory filters and excitation patters as representations of frequency resolution.* Academic, London, 1968.

[41] R. F. Lyon. A computational model of filtering, detection and compression in the cochlea. In *in Proceedings of International Acoustics Speech and Signal Processing, (IASSP'82)*, 1982.

[42] R. D. Patterson, M. Allerhand, and C. Giguere. Time-domain modelling of peripheral auditory processing: A modular architecture and software platform. *Journal of the Acoustical Society of America*, 98:1890–1894, 1995.

[43] S. Seneff. A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, 16, 1988.

[44] F. Baumgartner. *Ein psychophysiologisches Gehörmodell zur Nachbildung von Wahrnehmungsschwellen für die Audiocodierung.* PhD thesis, University of Hannover, Germany, 2002.

[45] T. Harczos, F. Klefenz, and A. Kátai. A neurobiologically inspired vowel recognizer using hough-transform. *International Conference on Computer Vision Theory and Applications*, 2006.

[46] R. Patterson, I. Nummo-Smith, and J. Holdsworth. An efficient auditory filterbank based on the gammatone function. *in Proceedings of the Institute of Acoustic Speecg Group on Auditory Modeling*, 1987.

[47] N. Deo and K. Grosh. Simplified nonlinear outer hair cell models. *Acoustical Society of America Journal*, 117:2141–2146, apr 2005.

[48] M. J. Hewitt and R. Meddis. Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse. *Journal of the Acoustical Society of America*, 87:1813–1816, 1990.

[49] C.J. Sumner, L.P. O'Mard, and E.A. Lopez-Poveda. A revised model of the inner-hair cell and auditory nerve complex. *Journal of the Acoustical Society of America*, 111:2178–2189, 2002.

[50] F. L. Wightman. The patterntransformation model of pitch. *The Journal of the Acoustical Society of America*, 54:407–416, 1973.

[51] E. Terhardt. Pitch, consonance and harmony. *The Journal of the Acoustical Society of America*, 55:1061–1069, 1974.

[52] W. A. Yost. *Fundamentals of Hearing: An Introduction*. Academic Press, London, 2000.

[53] R. Meddisa and M.J. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *The Journal of the Acoustical Society of America*, 89:2866–2882, 1991.

[54] R. Meddis and L. O. Mard. A unitary model of pitch perception. *The Journal of the Acoustical Society of America*, 102:1811–1820, 1997.

[55] S. Shamma and D. Klein. The case of the missing pitch templates: How harmonic templates emerge in the early auditory system. *The Journal of the Acoustical Society of America*, 107:2631–2644, 2000.

[56] L Wiegrebe and R. Meddis. The representation of periodic sounds in simulated sustained chopper units of the ventral cochlear nucleus. *The Journal of the Acoustical Society of America*, 115:1207–1218, 2004.

[57] B. Kollmeier and R. Koch. Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *The Journal of the Acoustical Society of America*, 95:1593–1602, 1994.

[58] WS Woods, M. Hansen, T. Wittkop, and B. Kollmeier. Using multiple cues for sound source separation. *Psychoacoustics, Speech and Hearing Aids*, 1995.

[59] L. O. Chua and L. Yang. Cellular Neural Networks: Theory. *IEEE Transactions on Circuits and Systems*, 35:1257–1272, 1988.

[60] L. O. Chua and L. Yang. Cellular Neural Networks: Applications. *IEEE Transactions on Circuits and Systems*, 35:1273–1290, 1988.

[61] T. Roska and L. O. Chua. The CNN Universal Machine: an Analogic Array Computer. *IEEE Transactions on Circuits and Systems-II*, 40:163–173, 1993.

[62] T. Roska. Computational and computer complexity of analogic cellular wave computers. *in Proceedings of the 7th IEEE International Workshop on Cellular Neural Networks and Their Applications, (CNNA 2002)*, pages 323–338, 2002.

[63] M. Brendel. *Two studies about the adaptivity of the Cellular Neural Networks.* PhD thesis, Analogical and Neural Computing Systems Laboratory, Computer and Automation Institute, Hungarian Academy of Sciences, Budapest, Hungary, 2001.

[64] J. M. Cruz and L. O. Chua. A CNN Chip for connected component detection. *IEEE Transactions on Circuits and Systems*, 38:812–817, 1991.

[65] H. Harrer, J. A. Nossek, and R. Stelzl. An analog implementation of Discrete-Time Cellular Neural Networks. *IEEE Transactions on Neural Networks*, 3:466–476, 1992.

[66] A. Rodríguez-Vázquez, S. Espejo, R. Domínguez-Castro, J. L. Huertas, and E. Sánchez-Sinencio. Current-mode techniques for the implementation of Continuous- and Discrete-Time Cellular Neural Networks. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 40:132–146, 1993.

[67] H. Harrer, J. A. Nossek, T. Roska, and L. O. Chua. A current-Mode DTCNN Universal Chip. *in Proceedings of the IEEE International Symposium on Circuits and Systems*, 4:135–138, 1994.

[68] R. Domínguez-Castro, S. Espejo, A. Rodríguez-Vázquez, and R. Carmona. A CNN Universal Chip in CMOS technology. In *in Proceedings of the IEEE International Workshop on Cellular Neural Networks and their Applications, (CNNA'94)*, pages 91–96, Rome, 1994.

[69] J. M. Cruz, L. O. Chua, and T. Roska. A fast, complex and efficient test implementation of the CNN Universal Machine. In *in Proceedings of the IEEE International Workshop on Cellular Neural Networks and their Applications, (CNNA'94)*, pages 61–66, Rome, 1994.

[70] A. Zarándy and Cs. Rekeczky. Bi-i: a standalone cellular vision system, part I. architecture and ultra high frame rate processing examples. In *in Proceedings of the Eight International Workshop on Cellular Neural Networks and their Applications, (CNNA04)*, pages 4–9, Budapest, 2004.

[71] S. Espejo, R. Dominguez-Castro, G. Linan, and A. Rodriguez-Vázquez. A 64x64 CNN Universal Chip with analog and digital I/O. In *in the Proceedings of 5th IEEE International Conference on Electronics, Circuits and Systems, (ICECS'98)*, pages 203–206, 98.

[72] G. Linan, A. Rodriguez-Vazquez, S. Espejo, and R. Dominguez-Castro. ACE16K: A 128x128 focal plane analog processor with digital I/O. In *in Proceedings of the seventh IEEE International Workshop on Cellular Neural Networks (CNNA2002)*, pages 132–139, 2002.

[73] Sz. Tokés, L. Orzó, A. Ayoub, and T. Roska. Laptop poac: A compact optical implementation of cnnum. In *in Proceedings of the Eight International Workshop on Cellular Neural Networks and their Applications, (CNNA04)*, pages 70–75, Budapest, 2004.

[74] I Petrás. *Spatio-Temporal Patterns and Active Wave Computing*. PhD thesis, Pazmany Peter Catholic University, Budapest, Hugary, 2005.

[75] T. Roska, L. Kék, L. Nemes, and Á. Zarándy. Cnn software library (templates, subroutines, and algorithms) version 8.1, 1999.

[76] Analogic Computers Ltd. The aladdin system.

[77] I. Szatmari, P. Foldesy, Cs. Rekeczky, and A. Zarandy. Image processing library for the Aladdin Visual Computer. *in Proceedings of the 7th IEEE International Workshop on Cellular Neural Networks and Their Applications, (CNNA 2002)*, pages 563–570, 2002.

[78] L. O. Chua and T. Roska. Cellular Neural Networks: Foundations and Primer. *Lecure Notes for the course EE129 at U. C. Berklet*, 1.7, 1998.

[79] P. Földesy. Statistical error modeling of CNN-UM architectures: the binary case. In *in Proceedings of the 7th IEEE International Workshop onCellular Neural Networks and Their Applications,(CNNA 2002)*, pages 467–474, 2002.

[80] P. Földesy. Statistical error modeling of CNN-UM architectures: the grayscale case. *World Scientific*, 2002.

[81] G. Linan, P. Foldesy, A. Rodrignez-Vazquez, and S. Espejoand R. Dominguez-Castro. Implementation of non-linear templates using a decompositiontechnique by a 0.5 μm CMOS CNN universal chip. In *in Proceedings of the IEEE International Symposium on Circuits and Systems, (ISCAS 2000)*, volume 2, Geneva, Italy, 2000.

[82] L. Kék. *CNN template dekompozíció - analogikai algoritmusok CNN-UM chip implementációjának egy lehetséges eszköze*. PhD thesis, Analogical and Neural Computing Systems Laboratory, Computer and Automation Institute, Hungarian Academy of Sciences, Budapest, Hungary, 1998.

[83] S. Malcom. Lyon.s cochlear model. Technical report, Apple Computer Ltd., 1988.

[84] K. R. Crounse and L. O. Chua. Arbitrary Spatial Convolution via CNN Universal Machine with 3x3 Templates: Methods and Issues. Technical Report UCB/ERL M96/5, University of Berkley, 1996.

[85] Analogic Computers Ltd. Instantvision eye-ris.

[86] S. T. Birchfield and D. K. Gillmor. Fast bayesian acoustic localization. *In Proceesings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP02)*, pages 1793–1796, 2002.

[87] T. Tarnóczy. *Akusztika : fizikai akusztika*. Akadémiai kiadó, Budapest, 1963.

[88] T. Tarnóczy. *Teremakusztika. I. Visszhangok és utózengés*. Akadémiai kiadó, Budapest, 1986.

[89] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, 65(4):943950, 1979.

[90] J. Borish. Extension of the image model to arbitrary polyhedra. *Journal of the Acoustical Society of America*, 75(6):1827–1836, 1984.

[91] U.R. Krockstadt. Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibrations*, 8(18), 1968.

[92] P. Heckbert and P. Hanrahan. Beam tracing polygonal objects. In *in Proceedings of International Conference on ACM Computer Graphics, (SIGGRAPH84)*, 119-127.

[93] T. A. Funkhouser, I. Carlbom, G. Elko, G. Pingali, M. Sondhi, and J. West. A beam tracing approach to acoustic modeling for interactive virtual environments. *in Proceedings of International Conference on ACM Computer Graphics, (SIGGRAPH84)*, pages 21–32, 1998.

[94] L. L. Beranek. *Concert and opera halls: how they sound*. American Institute of Physics, 1996.

[95] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, New York, NY, USA, 2001.

[96] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320327, 1976.

[97] S. M. Griebel and M. S. Brandstein. Microphone array source localization using realizable delay vectors. In *in Proceedings of IEEE Workshop of Applications of Signal Processing to Audio and Acoustics, (ASSP01)*, 2001.

[98] M. Brandstein, J. E. Adcock, and H. Silverman. A practical time-delay estimator for localizing speech sources with a microphone array. *Computer Speech and Language*, 9(2):153–169, 1995.

[99] M. Brandstein, J. Adcock, and H. Silverman. A closed-form location estimator for use with room environment microphone arrays. *IEEE Transactions on Speech and Audio Processing*, 5:45–60, 1997.

[100] A. Stéphenne and B. Champagne. A new cepstral prefiltering technique for estimating time delay under reverberant conditions. *Signal Processing*, 59(3):253–266, 1997.

[101] P. Svaizer, M. Matassoni, and M. Omologo. Acoustic source location in three-dimensional space using crosspower spectrum phase. In *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP97)*, pages 231–234, 1997.

[102] D. V. Rabinkin. *Placement for Microphone Arrays*. PhD thesis, New Brunswick, New Jersey, 1998.

[103] Krishnaraj Varma. *Time-Delay-Estimate Based Direction-of-Arrival Estimation for Speech in Reverberant Environments*. PhD thesis, Virginia Polytechnic Institute and State University, 2002.

[104] M. S. Brandstein. A pitch-based approach to time-delay estimation of reverberant speech. In *in Proceedings of IEEE Workshop of Applications of Signal Processing to Audio and Acoustics, (ASSP 97)*, 1997.

[105] J. Griffitsh W. Bangs. *Signal Processing*. Academic Press, 1973.

[106] G. Carter. Variance bounds for passively locating an acoustic source with a symetric line array. *Journal of Acoustic Society of America*, 62:922–926, 1977.

[107] W. Hanh and S. Tretter. Optimum processing for delay-vector estimation in passive signal arrays. *IEEE transaction on Information Theory*, 19:608–614, 1973.

[108] W. Hanh. Optimum signal processing for passive sonar range and bearing estimation. *Journal of Acoustical Society of America*, 58:201–207, 1975.

[109] M. Max and T Kailath. Optimal localization of multiple sources by passive arrays. *IEEE Transaction on Acoustc, Speech and Signal Processing*, 31:1210–1217, 1983.

[110] D.N. Zotkin and R. Duraiswami. Accelerated speech source localization via a hierarchical search of steered response power. *IEEE Transactions on Speech and Audio Processing*, 12(5):499–508, 2004.

[111] D.Ward, E. Lehmann, and R.Williamson. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Transactions on Speech and Audio Processing*, 11:826–836, 2003.

[112] V. M. Alvarado. *Talker Llocalization and Optimal Placement of Microphones for Linear Microphone Arrays using Stohastic Region Contraction.* PhD thesis, Brown University, 1990.

[113] E. E. Jan. *Processing of Large Scale Microphone Arrays for Sound Capture.* PhD thesis, Rutgers University, New Brunswick, NJ, 1995.

[114] R. J. Renomeron, D. V. Rabinkin, J. C. French, and J. L. Flanagan. Small-scale matched filter array processing for spatially selective sound capture. *134th Meeting of the Acoustical Society of America*, 102:3208, 1997.

[115] H.F. Silverman, W.R. Patterson, J.L. Flanagan, and D.V. Rabinkin. A digital processing system for source location and sound capture by large microphone arrays. In *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP97)*, 1997.

[116] D. H. Johnson and D.E. Dudgeon. *Array Signal Processing: Concepts and Techniques.* Simon & Schuster, 1992.

[117] S. Haykin. *Adaptive filter theory.* Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1996.

[118] R.O. Schmidt. *A signal subspace approach to multiple emitter location and spectral estimation.* PhD thesis, Stanford University, 1981.

[119] J. Krolik and D. Swingler. Focused wide-band array processing by spatial resampling. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(2):356–360, 1990.

[120] H. Wang and M. Kaveh. Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(4):823–831, 1985.

[121] K. M. Buckley and L. J. Griffiths. Broad-band signal-subspace spatial-spectrum (BASS-ALE) estimation. *EEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):953–964, 1988.

[122] S.T. Birchfield and D. K. Gillmor. Acoustic source direction by hemisphere sampling. In *In the Proceedongs of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. (ICASSP'01)*, volume 5, 2001.

[123] S. T. Birchfield. A unifying framework for acoustic localization. In *in Proceedings of the 12th European Signal Processing Conference, (EUSIPCO04)*, 2004.

[124] W.T. Chu and A. C. C. Warnock. Detailed directivity of sound fields around human talkers. Technical report, IRC Research Report 104, 2002.

[125] Catt-acoustic. http://www.catt.se.

[126] Odeon room acoustic. http://www.odeon.dk.

[127] J. Benesty. Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. *Journal of the Acoustical Society of America10*, 107:384–391, 200.

[128] M. Brandstein and H. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP97)*, pages 357–378, 1997.

[129] F. Talantzis, A. G. Constantinides, and L. C. Polymenakos. Estimation of direction of arrival using information theory. *IEEE Signal Processing Letters*, 12(8):561– 564, 2005.

[130] Y. Rui and D. Florencio. Time delay estimation in the presence of correlated noise and reverberation. In *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'04)*, volume 2, 2004.

[131] S. Doclo and M. Moonen. Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments. *EURASIP Journal on Applied Signal Processing*, 1110-1124:11, 2003.

[132] A. Epstein, G. U. Paul, B. Vettermann, C. Boulin, and F. Klefenz. A Parallel Systolic Array ASIC for Real-Time Execution of the Hough Transform. *IEEE TRANSACTIONS ON NUCLEAR SCIENCE*, 49(2):339, 2002.

[133] Y. A. Huang and J. Benesty. A class of frequency-domain adaptive approaches to blind multichannel identification. *IEEE Transactions on Signal Processing*, 51:11–24, 2003.

[134] G. Carter. Variance bounds for passively location an acoustic source with a symmetric line array. *Journal of the Acoustical Society of America*, 62:922–926, 1977.

[135] B. Ward, E. A. Lehmann, and R. C. Williamson. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Transactions on Speech and Audio Processing*, 11(6):826–836, 2003.

[136] J. P. Ianniello. Time delay estimation via cross-correlation in the presence of large estimation errors. *IEEE Transactions on Signal Processing*, 30:998–1003, 1982.

[137] L. E. Kinsler and A. R. Frey. *Fundamentals of Acoustics*. New York: John Wiley & Sons, 1962.

[138] L. Karlen. *Akustik i rum och byggander*. Svensk Byggtjänst, 1983. Svéd nyelven.

[139] N. Tsingos, I. Carlbom, G. Elko, R. Kubli, and T. Funkhouser. Validating acoustical simulations in the bell labs box. *IEEE Computer Graphics and Applications*, 22(4):28–37, 2002.

[140] M. Kleiner, R. Orlowski, and J. Kirszenstein. A comparison between results from a physical scale model and a computer image source model for architectural acoustics. *Applied acoustic*, 38:245–265, 1993.

[141] Personal Conversation with Bengt-Inge Dalenback. http://www.catt.dk.

Függelék

Average and threshold

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 0 \end{bmatrix} B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} z = -1.2$$

Déli lejtő detektor

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix} z = -0.3$$

Dilatáció

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} z = 8$$

Dilation left

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix} z = 5$$

Dilation right

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} z = 5$$

Edge e1

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} 0 & -1 & -1 \\ 1 & 1 & -1 \\ 0 & -1 & -1 \end{bmatrix} z = -6.5$$

Edge e2

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & -1 & -1 \end{bmatrix} z = -5.5$$

Edge e3

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} 0 & -1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & -1 \end{bmatrix} z = -4.5$$

Edge w1

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} -1 & -1 & 0 \\ -1 & 1 & 1 \\ -1 & -1 & 0 \end{bmatrix} z = -6.5$$

Edge w2

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} -1 & -1 & 1 \\ -1 & 1 & 0 \\ -1 & -1 & 0 \end{bmatrix} z = -6.5$$

Edge w3

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} -1 & -1 & 0 \\ -1 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix} z = -6.5$$

Északi lejtő detektor

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} 0 & -2 & 0 \\ 0 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix} z = -0.3$$

Masked shadow n

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1.8 & 0 \\ 0 & 1.5 & 0 \end{bmatrix} B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1.2 & 0 \\ 0 & 0 & 0 \end{bmatrix} z = 0$$

Masked shadow s

$$A = \begin{bmatrix} 0 & 1.5 & 0 \\ 0 & 1.8 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1.2 & 0 \\ 0 & 0 & 0 \end{bmatrix} z = 0$$

Match 2 sötét pixel 3x3-as régióban

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} z = 0$$

Recall

$$A = \begin{bmatrix} 0.5 & 0.5 & 0.5 \\ 0.5 & 4 & 0.5 \\ 0.5 & 0.5 & 0.5 \end{bmatrix} B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{bmatrix} z = 3$$

Shift east

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} z = 0$$

Shift south

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} z = 0$$

Threshold

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} z = -0.3$$

Vertical dilation

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} B = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} z = 2$$

Index