Pszichoakusztika és teremakusztika hangforrások tér-időbeni szimulált szegregációjában

*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

1. 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.
A több évtizedes kutatómunka eredményeként ma már léteznek olyan algoritmusok, melyek néhány százalékos hibával, az emberi hallgatók teljesítményéhez mérhetően képesek zaj és visszhangmentes körülmények között rögzített felvételek alapján beszédfelismerést végezni. A mesterséges rendszerek hatékonysága azonban az utcáról beszűrődő hangokkal, zenével, illetve egyéb háttérzajokkal terhelt beszéd felismerése esetén exponenciálisan romlik, míg azonos körülmények között az emberi hallgatók teljesítménye lényegében változatlan marad.
Környezetünket, szélsőséges helyzetektől eltekintve, tárgyak és azokkal kapcsolatos események formájában észleljük. Szimbolikus gondolkozásunk a valóságot is ilyen formán írja le, sőt a beérkező szenzoros információkat is forrásuk szerint rendezzük, azaz a környezetünkből érkező hangokat szétválasztjuk, csoportosítjuk, források szerinti szegregáljuk, a koktél-parti problémaként elhíresült feladatot oldjuk meg.
A hangforrások, hangobjektumok[3] aznosítása, térbeli és időbeni szétválasztása azonban mesterséges rendszerek számára eddig jobbára megoldhatatlannak bizonyult, mivel a tulajdonképpeni források által kibocsátott jelek elegyének szétválasztásán túl, a feladatot tovább bonyolítja a kibocsátott hanghullámokat visszaverő, torzító, fókuszáló, akusztikus környezetünket alkotó tárgyak hatása is. A környezetünkben fellelhető hangforrások térbeli és időbeni szegregációjának problematikáját vizsgálom. Megkülönböztetem a hangok fizikai jellemzők szerinti, heurisztikus algoritmusok szerinti csoportosítását, valamint külön foglalkozom az így azonosított hangobjektumok térbeli helyzetének meghatározásával.
Az első téziscsoportban a természetben előforduló hangkeltő fizikai folyamatok sajátosságait kiaknázó, az emberi hallórendszerben, pszichoakusztikus kísérletekkel igazolt csoportosítási algoritmusok Celluláris Hullámszámítógépen [1] való implementációját ismertetem. A második téziscsoport a különböző hangobjektumok térbeli elhelyezkedését meghatározó algoritmust mutatok be, mely az akusztikus környezet - a hangot visszaverő, illetve elnyelő felületek, helye, iránya - hatásait integrálva képes visszhangos környezetben elhelyezett anizotróp források helyének meghatározására.

2. Módszerek, eszközök

A kidolgozott 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

[1] A Binac volt az első számítógép, melyre a lyukkártya-olvasón kívül manuálisan is lehetett adatot rögzíteni egy, a géphez csatlakoztatott írógép-billentyűzet segítségével [http://inventors.about.com]
[2] Az egér kifejlesztője[http://inventors.about.com]
[3] A hallgató által azonos forrásból származóként kezelt ingerek csoportja

szimulátort. A szimulátorral előállított kétdimenziós spektro-temporális képfolyamon alkalmaztam a *hallási jelenet elemzés*[4] [2] 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 Celluláris Neurális Hálózatok (CNN), illetve a CNN Univerzális Gép (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.

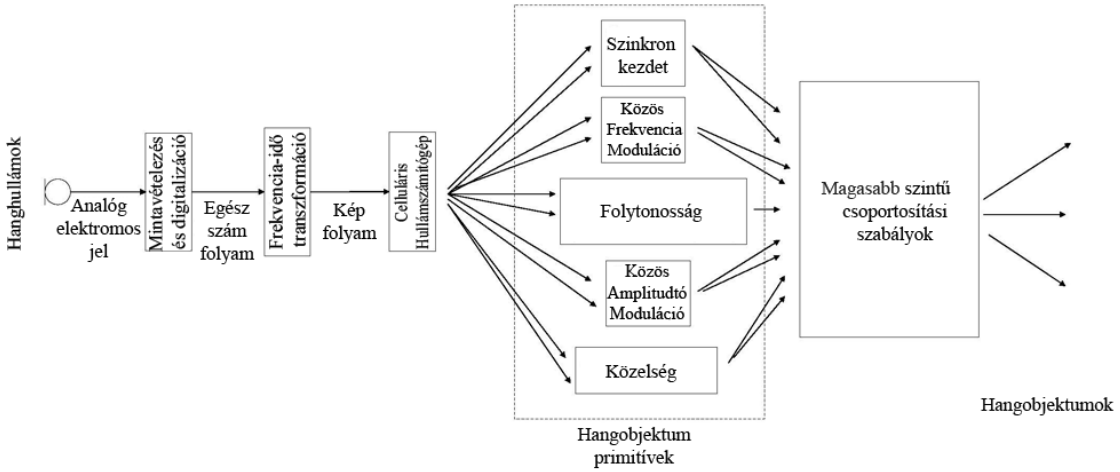
A hangforrások térbeli szegregációjának és elhelyezkedésének vizsgálatához a hang geometriai terjedésén alapuló modellt definiáltam. Tanulmányoztam a modell érvényességének határait, majd a matematikai analízis és a jelfeldolgozás eszközeit felhasználva következtetéseket fogalmaztam meg visszhangos környezetben 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 módszer teljesítményét C++-ban implementált rutinok segítségével a CAT akusztikus modellező szoftvert felhasználva ellenőriztem.

3. Új tudományos eredmények

1. Téziscsoport
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.

A *hallási jelenet elemzés* elmélete alapján [2] a hangokat olyan heurisztikus algoritmusokkal bontjuk hangobjektumokra, melyek az evolúció során a környezethez, a hangok fizikai természetéből adódó sajátosságokhoz alkalmazkodtak. A tárgyalt algoritmusokra jellemző, hogy nem emelhető ki olyan eljárás, ami mindig helyes eredményre vezet, ezért párhuzamosan több szempont szerinti kiértékelésére van szükség. Az általunk érzékelt hangobjektumok ezen algoritmusok eredményeként állnak elő. A közölt könyvtárban azoknak a csoportosítási algoritmusoknak Celluláris Hullámszámítógépen megvalósított funkcionális modelljei szerepelnek, melyek alapvető, velünk született „primitív" hang-szervezési formákat valósítanak meg. A csoportosításai algoritmusok elemi lépései - kétdimenziós spektro-temporális folyamon végzett műveletek - a hullámszámítógépek számára könnyen megoldható feladatok, ezért a könyvtár elemei valós időben párhuzamosan futtathatóak, így téve

[4] Auditory Scene Analysis



1. ábra. A *hallási jelenet elemzés* hullámszámítási architektúrája.

lehetővé az egyes algoritmusok kimenetének összevetésén alapuló döntést.

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 mintájára. 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 idő-

ben 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. Téziscsoport
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 hangterjedésmodell 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.

Akusztikus források helyének meghatározása a 1970-es évek óta aktívan kutatott terület [3]. Ekkor vált világossá, hogy a radar technikából ismert algoritmusok alkalmatlanok visszhangos körülmények közt elhelyezett széles spektrumú források helyének meghatározására. Az azóta eltelt mintegy harminc évben több kísérlet történt a probléma megoldására, azonban egyik sem vezetett kielégítő eredményre. Az általánosan elfogadott értelmezés szerint a visszaverődések által létrehozott „hamis" korrelációs csúcsok időbeni egybeesése okozza a módszerek megbízhatatlanságát. A disszertációban rámutatok, hogy a visszaverődések időbeni egybeesése nem szükséges feltétele a hibás eredmények létrejöttének, mivel a forrás iránykarakterisztikájából következően a visszavert hanghullámok csillapítása kisebb lehet a mikrofont és a forrást összekötő szakasz mentén terjedő hanghullám csillapításánál.

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*

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 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, illetve az eddigieknél zajmentesebb hangfelvételek készítésének lehetőséget eredményezi.

5. A további kutatás lehetséges irányai

Az 1. téziscsoportban bemutatott eljárások az emberi hallórendszernek csupán az adatvezérelt csoportosítási algoritmusainak megvalósításai. Az érzékelésben azonban 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[5] 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.
Mint az a 2. téziscsoportban közölt eredményekből következik, 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 dolgozatban e hatások integrációjára mutattam példát. A módszer meglevő zajérzékenységét kiküszöbölendő a jövőben érdemes megvizsgálni a visszhanghatások globális paraméterek alapján való figyelembevételének lehetőségeit. Tovább szélesítheti az algoritmus alkalmazási lehetőségeit a visszhanghatás-becslések több frekvenciatartományra való elkészítése, ami lehetővé teszi a rögzített jel spektrális tartalmához jobban illeszkedő becslések kiválasztását.
A szigorúan vett jelfeldolgozásnál messzebb vezet annak vizsgálata, hogy az élőlények testtartásának, illetve fejállásának az 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-időkülönbséget becslő algoritmusoknak.

6. 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.

[5] event related potencial

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.

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

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

Hivatkozások

[1] 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.

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

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