A HuComTech-korpusz és -adatbázis számítógépes feldolgozási lehetőségei. Automatikus prozódiai annotáció

Szekrényes István[1], Csipkés László[1], Oravecz Csaba[2]

[1] Debreceni Egyetem, Általános és Alkalmazott Nyelvészeti Tanszék
H-4032, Debrecen, Egyetem tér 1.
xepenator@gmail.com, laszlo.csipkes@freemail.hu
[2] Magyar Tudományos Akadémia, Nyelvtudományi Intézet
H-1394, Budapest, Pf. 360
oravecz@nytud.hu

Kivonat: A különböző kommunikációs események számítógépes elemzése során nélkülözhetetlen támpontot jelent, hogy gépileg feldolgozható formában elérhetők legyenek az azokat kísérő és általánosságban jellemző fizikai jegyek, mint amilyen a gyorsuló beszédtempó vagy az eltérő hanghordozás. A jelen tanulmányban bemutatásra kerülő, a HuComTech-korpusz és -adatbázis bővítéseként tervezett automatikus prozódiai annotáció ezeknek az információknak a feltérképezését szolgálja abból a célból, hogy a lehetővé tegye a korpusz annotációiban rögzítésre került kommunikációs jelenségek akusztikai jellemzését. A tanulmány a korpusz általános bemutatása után ennek céljait, módszereit és lehetőségeit kívánja részletezni.

1 Bevezetés

A HuComTech projekt[1] keretében létrehozott multimodális élőnyelvi korpusz és adatbázis számtalan feldolgozási és kutatási lehetőséget rejt magában. A kommunikációelméleti szakemberek, digitális képfeldolgozók és számítógépes nyelvészek közreműködésével, 113 beszélő részvételével gyűjtött, 50 órányi annotált anyag azzal a céllal készült, hogy egy egységes elméleti kerethez igazodva létrejöjjön egy olyan empirikus erőforrás, amely különféle kutatásokra, adatbányászatra, gépi betanításra alkalmas alapanyagot jelent a projektben együttműködő, illetve külső kutatók számára [4]. Jelen tanulmány a jelenlegi specifikációk rövid ismertetése után az adatbázis bővítéseként tervezett automatikus prozódiai annotációt, annak módszereit és lehetőségeit kívánja bemutatni.

[1] A kutatás alapjait *Az ember–gép kommunikáció technológiájának elméleti alapjai.* című, TÁMOP-4.2.2-08/1/2008-0009 projekt azonosítójú program keretei között teremtettük meg. Jelen tanulmány *A felsőoktatás minőségének javítása a kutatás-fejlesztés-innováció-oktatás fejlesztésén keresztül a Debreceni Egyetemen* című, TÁMOP-4.2.1/B-09/1/KONV-2010-0007 projektazonosítójú program keretein belül jött létre.

1.1 A HuComTech-korpusz és -adatbázis bemutatása

A HuComtech-korpusz magját egy összességében 50, beszélőnként fél óra hosszú audio- és videófelvétel alkotja. A felvételek mindegyike két személy (egy interjúztató és egy interjúalany) részvételével került rögzítésre, egy formális és egy informális társalgási szcenárió felhasználásával. Az első (formális) rész egy szimulált állásinterjú formájában, a második egy irányított beszélgetés szabadabb keretei között valósult meg, amelyek során az interjúztató különféle módszerekkel igyekezett az interjúalanyból spontán reakciókat kiváltani.

1. ábra: pillanatfelvétel a HuComTech korpuszból. Az interjúalany oldala.

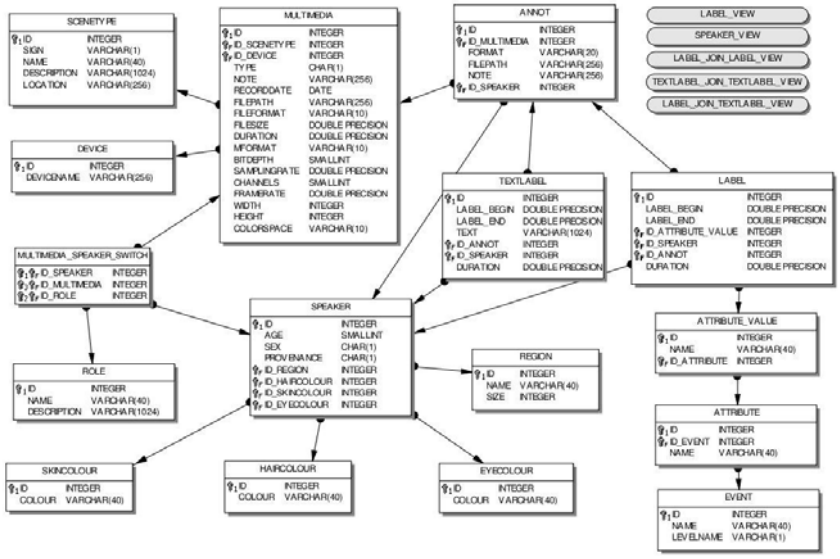
A korpusz számítógépes feldolgozhatóságát a felvételekhez készült annotációk biztosítják, amelyek elkészítésre az akusztikus és a vizuális csatornán párhuzamosan, többféle megközelítésben (fizikai jelek, nyelvi egységek és kommunikációs jelenségek megfigyelése), azokon belül is több elemzési szempont alapján történt.

A vizuális annotáció a képi anyagon megfigyelhető, a kommunikációs eseményeket kísérő, azok lehetséges jellemzőit képező fizikai jeleket rögzíti (fejmozgás, gesztikuláció, tekintetirány stb.), illetve interpretálja (arckifejezés jellege stb.). Az audioanyag szegmentálása során a beszédfolyam szintaktikai egységekre bomlik, amelyek mentén az annotáció a beszédfolyam szöveges átiratán kívül további információként tartalmazza annak hallás alapján meghatározott érzelmi töltését (a szemantikai tartalom figyelmen kívül hagyásával). Az így kinyerhető adatok a vizuális és akusztikus csatorna összefüggéseinek vizsgálatán túl a pragmatikai szempontú annotáció címkéivel összevetve válnak igazán informatívvá, ahol az annotátorok már nem nyelvi egységeket vagy fizikai jeleket, hanem kommunikációs eseményeket rögzítenek, vizuális, akusztikus és audiovizuális jegyek alapján.

Technológiai szempontból az audio- és a videócsatorna annotációja különböző számítógépes eszközökkel[2] és eltérő szegmentálási módszerekkel valósult meg, nem kizárva ezzel az utólagos konverziók, a modalitások egyesítése révén megvalósítható multimodális lekérdezéseket sem. Az annotációk tartalmazta adatok a feldolgozás során egy SQL-alapú adatbázis részeivé válnak, amely a felvételekkel kapcsolatos

[2] A videófelvételek rögzítésére a digitáliskép-feldolgozó csoport által fejlesztett Qannot, az audiofelvételek feldolgozására pedig a Praat beszédfeldolgozó szoftver szolgált [2].

különféle metainformációkat (beszélő neme, életkora stb.) is magában foglalja, az annotációs címkéket pedig a modellben elfoglalt helyük és tulajdonságtípusaik (arckifejezés, érzelmi töltés stb.) alapján rendszerezi (2. ábra).



2. ábra: A HuComTech adatbázisséma.

Az SQL lekérdezéseken kívül, a nyers adatokon (felvételek és annotációk) folytatott munka a feldolgozás azon részét képezi, amely egyúttal a korpusz bővítését is magával vonja az automatikusan generált új annotációk vagy metaadatok formájában. Az automatizált adatgyűjtés és címkézés ilyen számítógépes nyelvészeti irányú részét képezi a különféle akusztikai információk kinyerése és annotálása a már meglévő manuális annotációk felhasználásával.

1.2 Az automatikus prozódiai annotáció szerepe az adatbázisban

A prozódiai annotációval ellátott beszélt nyelvi korpuszok rendkívül értékes nyelvi erőforrást képviselnek, ám előállításuk igen munkaigényes. További problémát okoz, hogy a nemzetközi gyakorlatban nincs egyértelmű megállapodás arra vonatkozóan, hogy pontosan mit is tartalmazzon egy prozódiai annotáció.

Saját annotációs eljárásunk megtervezése során a távlati célok figyelembevételével azokat az elemzési megközelítéseket tekintettük megfelelőnek, amelyek az adatbázisban jelölésre került kommunikációs események gépi detektálásához szolgáltathatnak releváns információkat. Ennek megfelelően a kommunikációs eseményeket kísérő, általánosságban jellemző és valós időben is feldolgozható fizikai jegyeket szükséges

elemezhetővé tenni, amelyek együttesei, meghatározott irányú progressziója alapján amazok felismerhetővé válnak.

A pragmatikai annotációkban jelölt kommunikációs események ilyen értelemben vett potenciális kísérőjegyei vizuális oldalon részben manuálisan, részben automatikusan (pl. a szájmozgás) rögzítésre kerültek, detektálásuk pedig a digitális képfeldolgozás feladatkörébe esik, a kapcsolódó prozódiai információk viszont az adatbázis jelenlegi állapotában egyáltalán nem elérhetők. Az automatikus prozódiai annotáció célja pótolni ezt a hiányt, hogy a nyers adatok (F0 és intenzitásértékek) az adatbázisban közvetlenül, illetve a különféle címkézési eljárások révén feldolgozott formában is lekérdezhetővé váljanak. A feldolgozás eredményeként kapott címkesorokból aztán tágabb körű elemzések útján további metainformációk nyerhetők ki az interakciók beszéddinamikai mintázatairól, amelyek feltérképezése által a kommunikációs események felismerését segítő tudás birtokába juthatunk. Például arról, hogyan változik egy dialógus intenzitása az abba bekerülő új információk, témaváltások hatására.

2 A prozódiai annotáció lépései

2.1 F0- és intenzitásadatok kinyerése és integrálása az adatbázisba

A beszédfolyam akusztikai karakterizálásához leginkább felhasználható F0 és intenzitás adatok kinyerésére a Praat beszédfeldolgozó szoftver [2] e célra kidolgozott, beépített szkript nyelve által könnyen automatizálható lekérdező funkciói mellett saját fejlesztésű, valós időben is működő, jelenleg tesztelés alatt álló algoritmusokat kívánunk a későbbiekben felhasználni. Ezek tetszőleges formára hozható kimenete a korpusz részeként további elemzések bemenetéül szolgál, illetve feltöltésük után az eredmények az adatbázis-lekérdezések során is felhasználhatóvá válnak.

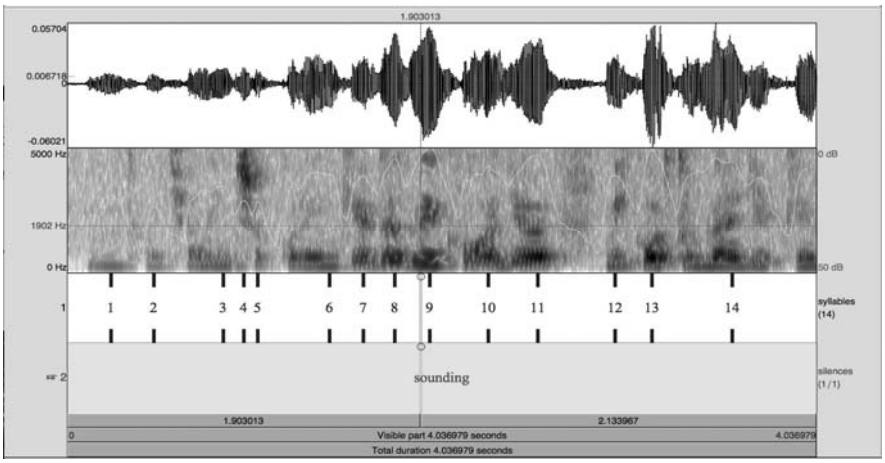
A HuComTech projekt jelenlegi adatbázissémája egyetlen relációs táblában tárolja a különböző típusú annotációk címkéit a címkekezdet, címkevég, címkeérték oszlopokban rögzítve az azokat jellemző legfontosabb információkat (lásd 1. ábra). Az olyan típusú akusztikai adatok, mint az egy adott időpillanathoz tartozó F0- és intenzitásértékek tárolására ez a tábla nem alkalmas, így a többi annotációs címkétől szeparáltan, külön táblában kerülnek tárolásra, amely később alkalmas egyéb, megegyező struktúrájú (idő → érték) fizikai adatok tárolására is. Ezek az adatok a lekérdezések során természetesen csak bizonyos kalkulációk, például bizonyos címkeszakaszokra vagy az egész fájlra számolt átlagértékek után válnak kellően informatívvá.

2.2 A beszédtempó annotációja

A feldolgozási eljárás egyik fontos komponensét a beszédtempó mérése és címkézése jelenti, melynek során a beszéd sebességének változásairól kívánunk számot adni.

A beszédtempó mérésének kivitelezéséhez elsősorban egy olyan mérési objektum meghatározására van szükségünk, amelynek egy adott időegységre mért gyakorisága,

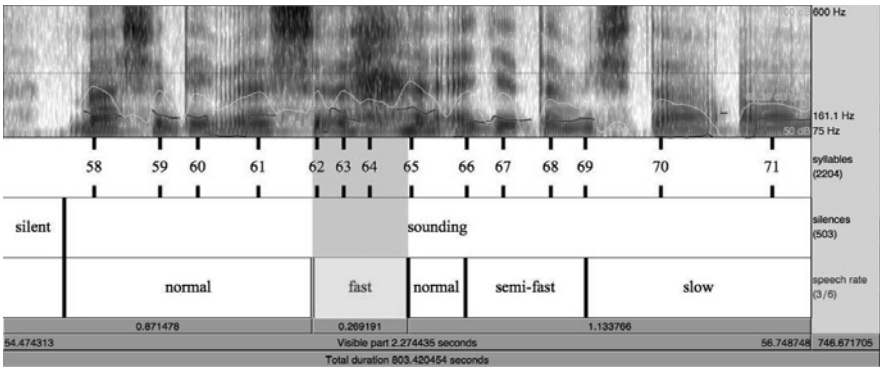
sűrűsége megragadhatóvá teszi azt. A létező megoldások után kutatva találtunk rá Nivja H. de Jong és Ton Wempe tanulmányára [3]. A szerzők a beszédtempó vizsgálatához a szótagmagokat választották mérési objektumként, amelyek detektálására egy jól működő módszert is kidolgoztak. Az eljárás Praat beszédfeldolgozó program beépített szkript nyelvét, függvényeit és mérési algoritmusait használja. A szótagmagok detektálása az intenzitás görbe csúcsainak meghatározott küszöbértékek (csúcsok közötti minimális értékbeli különbség stb.) szerinti szűrése által történik a beszédfolyam nem hangzós részeinek kizárásával. Az eredményül kapott intenzitáscsúcsok időpillanatai a Praat TextGrid formátumú annotációs fájljaiban kerülnek tárolásra, amelyek a program szerkesztőfelületén jeleníthetők meg (2. ábra), illetve egyéb szoftveres megoldásokkal is könnyen feldolgozhatók.



3. ábra: A szótagmagok detektálása.

A beszéd sebességének ingadozása így a szótagmagok helyét reprezentáló intenzitáscsúcsok közötti távolság változásain keresztül válik megragadhatóvá.[3] Ehhez természetesen figyelembe kell vennünk a beszéd sebességének az adott beszélő egyedi beszédtempójából következő relatív viszonyait, amely a teljes beszédfolyamra számolt előzetes statisztikák segítségével valósítható meg. A hangzós részekre számolt csúcsok közötti távolság átlagértékének megadásával meghatározhatjuk az adott beszélő normál beszédtempóját. Az eljárás során az átlagolást először minden hangzós szakaszra külön-külön végezzünk el, majd ezeket az eredményeket átlagoljuk újra. A normál beszédtempó meghatározása után relatív küszöbértékek kiszámításával további kategóriákat állíthatunk fel, amelyek már az adott szakaszokra történő címkézési eljárás során kerülnek felhasználásra (3. ábra).

[3] A különböző magánhangzók eltérő ejtési idejéből fakadóan ez az eljárás könnyen vezethet megtévesztő eredményekhez. Az algoritmus tökéletesítéséhez tehát plusz információként figyelembe kell venni a csúcsok által reprezentált szótagmag időbeli terjedelmét is, amely az F0- és az intenzitásgörbe további vizsgálata révén lesz megvalósítható.



4. ábra: A beszédtempó címkézése.

A beszéd aktuális tempóját tehát az adott szegmensen belül fellelt szótagmagok átlagsűrűségének az adott beszélőre jellemző normál átlagsűrűséghez viszonyított különbsége fogja meghatározni a beszéd aktuális tempóját. A eljárás lépéseit összefoglalva:

- szótagmagok detektálása (de Jong és Wempe munkája [3] nyomán)
- normál beszédtempó meghatározása a szótagmagok hangzós részekre számolt átlagsűrűsége alapján (beszélőspecifikus tulajdonság)
- az adott beszédszegmens átlagsűrűségének kiszámítása
- az adott beszédszegmens tempójának kategorizálása a normál beszédtempótól való eltérés foka alapján

A címkézés esetében problematikus kérdés, hogy milyen egységekre, a beszédfolyam mely szakaszaira történjen az aktuális beszédtempó kategorizálása. Lehetséges utat jelent a korábban már manálisan annotált szegmentumok, illetve a szünettől szünetig tartó hangzós részek tempójának címkézése. Az eljárásnál problémát jelent, hogy egy folytonos (szünettől szünetig tartó) beszédszakaszon, vagy akár egy szintaktikai egységet reprezentáló annotált szegmentumon belül is számítanunk kell a tempó ingadozására. Hogy ezeket az információkat ne veszítsük el, az adott egységen belül is vizsgálnunk a beszédtempó alakulását, a beszélőt és az egységet jellemző adatokból számolt küszöbértékek felhasználásával.

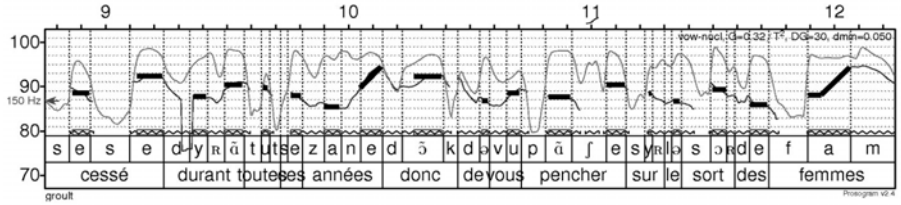
2. 3 Az alapfrekvencia progressziójának annotálása

A prozódiai annotáció következő lépését az alapfrekvencia progressziójának elemzése jelenti, amelynek eredményeként a beszédfolyam meghatározott szegmentumaihoz valamilyen egzakt tonális karaktert jelölő annotációs címkét (emelkedő, ereszkedő, eső stb.) vagy címkekombinációt rendelünk. Ennek megvalósítása érdekében a kimért F0-értékekre számolt trendvonalak formájában előbb feldolgozható formában stilizálnunk kell az alapfrekvencia változásait.

Az eljárás megvalósítására Piet Mertens kapcsolódó munkáját [5] terveztük felhasználni. Mertens előzetesen számos fontos feltételt fogalmaz meg, amelyeket a prozódiai annotáció során nem szabad figyelmen kívül hagyni:

- Az annotácónak alapvetően az érzékelhető intonációt kell reprezentálnia ojektív és könnyen értelmezhető módon,
- Az alapfrekvencia változását hosszabb beszédfolyamon keresztül is tükröznie kell, a szélesebb tartományokra kiterjedő változások rögzítése érdekében,
- A fizikai jelek időbeli szerveződését meg kell őriznie a szünetek, hezitációk, beszédtempó és a ritmus azonosíthatósága érdekében,
- Az annotációnak automatikusnak vagy félautomatikusnak kell lennie,
- Az annotáció elméletsemleges kell, hogy legyen, a széleskörű használhatóság érdekében,
- Az annotáció lehetőleg időben illesztett fonetikai és szöveges átírást tartalmazzon az olvashatóság és szöveges keresés lehetőségének biztosítása érdekében.

Mertens [5] kifejlesztett egy, a fenti feltételeknek megfelelő transkripciós rendszert, amely a vokális szótagmag alapfrekvenciájának stilizált kontúrját felhasználva félautomatikusan rendel prozódiai annotációt fonetikai transkripcióhoz. A stilizálás [1] alapján a tonális érzékelés pszichoakusztikai modelljére épül. Az annotáció megőrzi az akusztikai jel temporális jellemzőit, és beépíti a szöveges, illetve a fonetikai transkripciót is, ahol ez utóbbi a vokális szótagmag azonosításában játszik szerepet. A rendszer többféle részletességű információt tartalmazó kimenetet képes generálni: a kompakt változat a stilizált beszéddallam szöveges és fonetikai átírással kiegészített annotációját tartalmazza (lásd 5. ábra).



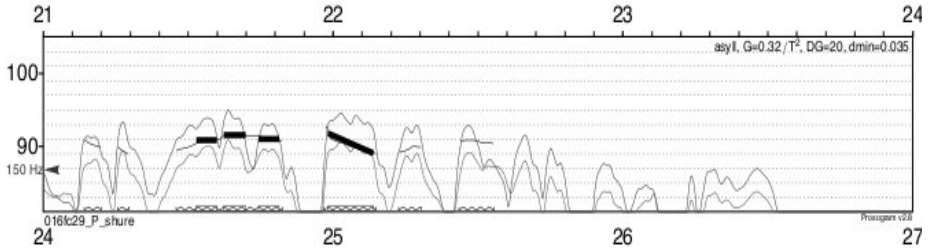
5. ábra: A Mertens-féle transkripciós rendszer kimenete.

A módszer implementációja a Praat beszédfeldolgozó program felhasználásával történt. A transkripciókat generáló Praat szkript a hozzá tartozó dokumentációval együtt Prosogram (v2.8) néven szabadon hozzáférhető[4], többféle beállítással és üzemmódban futtatható, lehetőséget biztosítva például meglévő, a megfelelő formátumban tárolt manuális szegmentációk használatára. A HuComTech-korpuszban hozzáférhető szöveges transkripciók tagmondatszintű annotációkat takarnak, így az alapfrekvencia félautomatikus stilizációjához ezek nem

[4] http://bach.arts.kuleuven.be/pmertens/prosogram/

felhasználhatók, viszont a program lehetőséget kínál a hanganyag szótagokra és szótagmagokra történő automatikus szegmentálására is.[5]

Az eredményül kapott stilizációknak[6] a felhasználásával további elemzésével lehetővé válik a beszédfolyam szegmentumainak egzakt kategorizációja. Problémát jelent viszont, hogy a stilizációkat tartalmazó kimenet csak grafikus formában elérhető. A általunk tervezett, a HuComTech adatbázisba integrálható prozódiai annotáció megvalósításához így a stilizációk megjelenítésért felelős algoritmust előbb vissza kell fejtenünk és át kell alakítanunk, hogy a célnak megfelelő, a további számításokhoz felhasználható numerikus kimeneteket (a stilizációk kezdő és végpontja) tudjunk produkálni. A program saját anyagunkon végzett tesztelésének grafikus kimenetét az 5. ábra szemlélteti.



6. ábra: A Prosogram grafikus kimenete.

A további elemzések bemenetét tehát az alapfrekvencia stilizált progressziója adja, amely a dallamgörbe normalizált darabjainak hosszában, a kezdő és végpontok frekvenciaértékének különbségében ragadható meg. Ezeknek az értékeknek a felhasználásával történik a beszédfolyam tonális egységeinek címkézése, ahol minden címke az adott egység dallamának karakteréről próbál feldolgozható leírást adni.

Mint ahogyan a beszédtempónál, az alapfrekvencia annotálásánál is problémát jelent, hogy a beszédfolyamnak melyek azok az egységei, amelyek kiértékelése révén az alapfrekvencia változásairól a számunkra megfelelő léptékű képet kapjuk. A jelenlegi tervek szerint ezek az egységek a korpuszban már manuálisan annotált, potenciális intonációs frázisokat jelentő tagmondatok lesznek, nem kizárva a dallammenet tágabb léptékű, különféle kommunikációs események mentén történő elemzését. Ezekhez a vizsgálatokhoz célszerű a tagmondatszintű progresszió kategorizálása mellett számot adni a beszéddallam aktuális tartományáról, annak relatív magasságának függvényében.[7]

[5] Ennek megbízhatósága saját anyagunkon jelenleg tesztelés alatt áll.

[6] Amelyeket a továbbiakban az alapfrekvencia normalizált progressziójának tekintünk.

[7] Ennek a relatív magasságnak a meghatározásához az adott beszélőre jellemző hangterjedelem szolgáltat információkat.

3 Összegzés

A HuComTech-korpusz és -adatbázis jelenlegi állapotában számos vizsgálati lehetőséget biztosít kommunikációelméleti kutatások folytatására. Az automatikus prozódiai annotáció sikeres implementációja jelentős mértékben kitágítja ezeket a vizsgálati lehetőségeket az akusztikai információk feldolgozható formában történő bekapcsolásával, olyan további kutatásokat alapozva meg, melyek egy adott kommunikációs esemény valós időben történő detektálásának vagy predikciójának algoritmizálhatóságát célozzák.

Bibliográfia

1. Alessandro, P., Mertens., P.: Automatic pitch contour stylization using a model of tonal perception. Computer Speech & Language Vol. 9, No. 3 (1995) 257-288
2. Boersma, P., Weenink, D. (2010): Praat: doing phonetics by computer 5.1.43. Institute of Phonetic Sciences, University of Amsterdam. http://www.praat.org
3. de Jong, N. H., Wempe, T.: Praat script to detect syllable nuclei and measure speech rate automatically. Behavior Research Methods Vol. 41, No. 2 (2009) 385-390.
4. Hunyadi, L.: Multimodal human– computer interaction technologies. Theoretical modeling and application in speech processing. Argumentum. Megjelenés alatt (2011)
5. Mertens, P.: The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model. In: Bel, B., Marlien, I. (eds.): Proceedings of Speech Prosody 2004i, Nara (Japan), 23-26 March (ISBN 2-9518233-1-2) (2004)
6. Pápay, K., Szeghalmy, Sz., Szekrényes, I.: HuComTech Multimodal Corpus Annotation. Argumentum. Megjelenés alatt (2011)