

Doktori (PhD) értekezés

**Prozódiai jellemzők gépi feldolgozása és
hasznosítása élőnyelvi korpuszok elemzésében**

Szekrényes István

Témavezető: Dr. Hunyadi László



DEBRECENI EGYETEM

Nyelvtudományok Doktori Iskola

Debrecen, 2019

PROZÓDIAI JELLEMZŐK GÉPI FELDOLGOZÁSA ÉS HASZNOSÍTÁSA ÉLŐNYELVI KORPUSZOK ELEMZÉSÉBEN

Értekezés a doktori (Ph.D.) fokozat megszerzése érdekében
a nyelvtudományok tudományágban

Írta: Szekrényes István okleveles elméleti nyelvész – filozófia szakos bölcész

Készült a Debreceni Egyetem Nyelvtudományok doktori iskolája
(Általános és Alkalmazott Nyelvészeti programja) keretében

Témavezető: Dr.
(olvasható aláírás)

A doktori szigorlati bizottság:

elnök: Dr.
tagok: Dr.
Dr.

A doktori szigorlat időpontja: 200... ..

Az értekezés bírálói:

Dr.
Dr.
Dr.

A bírálóbizottság:

elnök: Dr.
tagok: Dr.
Dr.
Dr.
Dr.

A nyilvános vita időpontja: 200... ..

Én, Szekrényes István, teljes felelősségem tudatában kijelentem, hogy a benyújtott értekezés önálló munka, a szerzői jog nemzetközi normáinak tiszteletben tartásával készült, a benne található irodalmi hivatkozások egyértelműek és teljesek. Nem állok doktori fokozat visszavonására irányuló eljárás alatt, illetve 5 éven belül nem vontak vissza tőlem odaítélt doktori fokozatot. Jelen értekezést korábban más intézményben nem nyújtottam be és azt nem utasították el.

(Szekrényes István)

Köszönetnyilvánítás

Ezúton szeretném kifejezni köszönetemet témavezetőmnek és tanáromnak, Prof. Dr. Hunyadi Lászlónak, aki egészen hallgató koromtól figyelemmel kíséri és segíti pályámat. Az Általános és Alkalmazott Nyelvészeti Tanszék vezetőjeként lehetőséget adott rá, hogy a szakmában kipróbáljam magam, a kutatással és a disszertáció írásával eltelt évek alatt pedig mindig készen állt arra, hogy meghallgasson, elolvasson, kijavítson, tanácsot adjon, vagy épp a következő lépésre ösztökéljen.

A disszertációval kapcsolatos konzultációkért, kritikákért, tanácsokért köszönettel tartozom Dr. Bódog Alexának, Dr. Kelemen Istvánnak, Prof. Dr. Kornai Andrásnak, Dr. Márton Sándornak, Dr. Tóth Enikőnek és Prof. Dr. Winkler Istvánnak.

Köszönettel tartozom opponenseimnek, Dr. Mády Katalinnak és Dr. Szaszák Györgynek, akik az értekezés munkahelyi vitára leadott verzióját rendkívül alaposan és konstruktívan véleményezték, ami a szöveg átdolgozásában és az elvégzett munka realisabb önértékelésében is nagy segítséget jelentett.

Köszönettel tartozom az e-magyar projekt csapatának, különösen Prof. Dr. Kornai Andrásnak, Dr. Simon Eszternek és Dr. Váradi Tamásnak, akik a munkámnak bizalmat szavazva a fejlesztéseimet befogadták és azóta is számos formában segítik, hogy a tágabb szakmai közösségekben is otthonosan érezzem magam. Prof. Dr. Kornai Andrásnak külön hálás vagyok azokért az új tapasztalatokért, amiket egyéb projektjeinkben szerezhettem.

A közös munkáért köszönettel tartozom Dr. Kovács Györgynek, aki társszerzőmként a neurális hálózatokkal végrehajtott kísérletekben nélkülözhetetlen segítséget jelentett.

Köszönettel tartozom Kovács László Eriknek, akitől egy közösen tartott szakkollégium során rengeteg hasznos visszajelzést és tanácsot kaptam a prozódiai annotáció vizuális kimeneténél és a hozzá tartozó értékelő felület kidolgozásához.

Köszönettel tartozom Ghazaleh Esfandiari-Baiat kollégámnak, akivel annak idején a témaváltás prozódia alapján való észlelhetőségét – a disszertáció egyik első, tapogatózó lépéseként – közösen tesztelni kezdtük.

Köszönettel tartozom egykori munkatársaimnak, akikkel az évek során valódi csapattá kovácsolódva hoztuk létre a HuComTech korpuszt. Név szerint: Dr. Abuczki Ágnesnek, Dr. Bódog Alexának, Borbély Anikónak, Csipkés Lászlónak, Földesi Andrásnak, Kiss Herminának, Oravecz Csabának, Pápay Kingának, Dr. Szeghalmy Szilviának, Takács Orsolyának és Tündik Máté Ákosnak. Köztük külön köszönettel tartozom Pápay Kingának, aki a HuComTech projekt kezdetén alprojekt vezetőként a szárnyai alá vett.

Köszönettel tartozom Dr. Bodnár János Kristófnak, Dézsi Éva Darinkának, Dr. Karap Zoltánnak, Dr. Péter Szabinának és Sas Szilárdnak a percepció kísérleteim elő-tesztelése során adott hasznos észrevételekért és a fejlesztéseim kimenetének értékeléséért.

Köszönettel tartozom Dr. Csátár Péternek és Dr. Valastyán Tamásnak az idő múlása és a disszertáció állapota közötti összefüggéseket jóindulatúan megvilágító biztatásért, a megértő türelemért és a sok segítségért, amivel egyéb módokon is segítettek a munkámat. Ugyanezért szeretném kifejezni köszönetemet a Filozófia Intézet munkatársai közül Gáspár Lászlónak, Dr. Horváth Lajosnak, Dr. Kelemen Istvánnak, Dr. Pólik Józsefnek, Szakályné Lantai Lillának és Dr. Tánczos Péternek.

Köszönettel tartozom Feketéné Balogh Mariannának és Dr. Szendrei Ákosnak, akik a fokozatszerzési eljárás szükséges lépéseiben rengeteget segítettek.

Végül, de nem utolsó sorban szeretnék köszönetet mondani családtagjaimnak és barátaimnak minden segítségért, támogatásért, türelemért és megértésért, amivel a munkámat igyekeztek megkönnyíteni. Kiemelt köszönettel tartozom szüleimnek, Angyal Orsolyának és családjának, továbbá Radák Balázsnak, akik segítségére és/vagy megértésére az utóbbi időszakban leginkább rá voltam szorulva.

Tartalomjegyzék

1. Bevezetés	9
1.1. A kutatás háttere	9
1.2. A kutatás célja	12
1.3. A disszertáció szerkezeti felépítése	13
2. A kutatás elméleti háttere	14
2.1. A prozódiai jellemzők természetéről	14
2.2. A prozódia helye a nyelv és a kommunikáció modelljeiben	17
2.2.1. „Színtelen zöld ideák intonálnak-e?”	17
2.2.2. Megnyilatkozások és beszédaktusok	21
2.3. A társalgás prozodiájáról	23
3. A kutatás beszédtechnológiai háttere	30
3.1. A beszéd detektálása és szegmentálása	30
3.2. Beszélő diarizáció	34
3.3. Az intonáció modellezése és lejegyzése	37
3.3.1. ToBI	38
3.3.2. Tilt	40
3.3.3. Prosogram	43
4. A kutatás anyaga és módszertani eredményei	46
4.1. A kutatás anyaga	46
4.1.1. A felvételek	48
4.1.2. A korpusz feldolgozásának folyamata	50
4.1.3. A korpusz annotációs sémái	52
4.2. A prozódia annotációjával kapcsolatos fejlesztések	54
4.2.1. A beszéddallam annotációja	56

4.2.1.1.	Manuális annotáció (kísérletek)	57
4.2.1.2.	Az algoritmus rövid leírása	60
4.2.1.3.	Előfeldolgozás és bemeneti kritériumok	63
4.2.1.4.	A beszélő egyedi hangterjedelmére vonatkozó küszöbértékek számítása	66
4.2.1.5.	Az F0-adatok feldolgozása	71
4.2.1.6.	A dallamformák stilizálása, szegmentálása és osztályozása .	77
4.2.2.	Az intenzitás annotációja	82
4.2.3.	A beszédtempó annotációja	83
4.2.4.	XML konverzió	87
4.2.5.	A kimenet vizualizálása XSLT transzformációval	89
4.2.6.	A kimenet értékelése	91
5.	Percepciós kísérletek	98
5.1.	Idegen fülek a Perzsa öbölben	98
5.1.1.	A kísérleti felvételek	99
5.1.2.	A kísérleti kérdések	102
5.1.3.	Eredmények	103
5.2.	Emelkedik vagy csökken?	104
5.2.1.	Mérési módszerek	105
5.2.2.	Kísérleti feladatok és kondíciók	105
5.2.3.	Az adatok feldolgozása	108
5.2.4.	Eredmények	112
5.3.	„Angéla néni” esete	115
5.4.	Formális vagy informális?	124
6.	Kísérletek gépi tanulással	127
6.1.	Formális és informális dialógusok megkülönböztetése	127
6.1.1.	Az osztályzás anyaga	128
6.1.2.	Felhasznált jellemzők	129
6.1.2.1.	Beszélőváltás	129
6.1.2.2.	Intonáció	133
6.1.3.	A neuronhálók tanítása és tesztelése	135
6.1.4.	A kísérlet eredményei	137

6.2. A témaváltások detektálása	138
7. Összefoglalás	144
Irodalomjegyzék	149
Melléklet	159
Az 5.3. fejezethez tartozó ábrák	160
Az értekezés magyar nyelvű összefoglalója	165
Az értekezés angol nyelvű összefoglalója	166
A PhD értekezés alapjául szolgáló közlemények	167

1. FEJEZET

Bevezetés

1.1. A kutatás háttere

A prozódia a „beszéd zenei kifejezőeszközeiként” aposztrofált és a „kommunikáció nem nyelvi elemei” közé besorolt jelenségei - az idézett fordulatok pejoratív értelmezésében - könnyen kelthetik azt a benyomást, hogy az információközlés absztraktabb nyelvi szintjeihez képest a nyelvhasználat afféle addicionális, esztétika-retorikai kellékeivel állunk szemben a nyelvtudomány egy kimondottan periferiális területén. Ha tehát a nyelvi jel elsődleges, konstatív, logika tartalmára koncentrálunk, a beszéd zenei szerveződése tisztán formai adalék gyanánt (mint „színezet”) ignorálható. Néhány, a nyelvtudomány szempontjából belső és külső tényezőnek köszönhetően azonban a prozódia fönt vázolt státusza napjainkra jelentősen megváltozott.

Az egyik fontos, „belsőnek” nevezhető tényező a beszédaktus-elmélet (Austin, 1962), a pragmatika performatív fordulatként is emlegetett diszciplínájának megszületése, melynek keretében a kommunikáció alapegységeit képező megnyilatkozások nem pusztán mint proposícionális tartalmak hordozói, de mint ún. illokúciós erővel rendelkező performatív aktusok kerültek a vizsgálódás fókuszába. Az illokúciós erők azonosítása és tipologizálása során (az egyéb, főként kontextuális információk mellett) a prozódia szerepe és jelentősége is felértékelődik.¹ Mindez nem jelenti azt, hogy a prozódia ne játszana fontos szerepet a nyelvi szerkezetek formális szemantikában vizsgált konvencionális jelentésének meghatározásában is. Számos olyan példát találhatunk, ahol a szerkezeti többértelműségek (logikai operátorok hatóköre, fókusz és topik pozíció stb.) csak a szupraszegmentumok (itt elsősorban a hangsúlyra és az intonáció-

¹ A prozódiai információk a megnyilatkozások modalitásának azonosításán túl, például a diskurzus-partikulák aktuális funkciójának meghatározásában is segítséget jelenthetnek (Abuczki, 2013).

ra gondolunk) disztinktív szerepének játékba hozásával oldható fel.² Habár a nyelvről alkotott generatív modellekben ez a modul a szintaxisához képest háttérbe szorul, léteznek olyan megközelítések, melyek a prozódiai szerkezet függetlenségét³ és a logikai forma megfelelő értelmezésében betöltött szerepét kimondottan hangsúlyozzák (Jackendoff, 2002; Hunyadi, 2002). Ennek az a jelentősége, hogy a prozódia nem csupán a nyelvhasználat pragmatikai oldaláról, de a grammatika inherens részeként járulhat hozzá a felszíni szerkezet megformálásához. Nem kell meglepődnünk tehát, hogy a magyar nyelvre formális alapon kidolgozott fonológiai rendszerekben a hangsúly és az intonáció előkelő helyet kapott (Kornai & Kálmán, 1988; Varga, 1994).

A vizsgálatok közegét érintő, további általános tendenciaként figyelhető meg a nyelvészek élénk és egyre növekvő érdeklődése a spontán beszéd kutatása iránt, amely egyfajta hangsúlyeltolódásként tartható számon a nyelvtudomány klasszikus médiumát jelentő *szövegtől* az *élőszó* felé.⁴ A spontán beszéd kutatásához létrehozott multimodális élőnyelvi korpuszok a kommunikáció non-verbális kifejezőeszközeinek vizsgálatának is kiemelt szerepet hagyva fontos empirikus erőforrást jelentenek ehhez.

„Külső” tényezőként könyvelhető el azoknak az új technológiáknak és módszereknek a megszületése és széles körű elterjedése, amelyek a beszéd akusztikai szerkezetét és szupraszegmentális alapelemeit (alapfrekvencia és intenzitás) számítógépes eszközökkel könnyen mérhetővé és elemezhetővé teszik. A nyelvészek prozódiával kapcsolatos, többnyire introspekción alapuló megfigyelései és hipotézisei ezáltal természettudományos módszerekkel verifikálhatók, bővíthetővé, interdiszciplináris kutatások számára is könnyebben hozzáférhetővé, a nyelvtechnológiában pedig alkalmazhatóvá váltak. A modern kommunikációs technológiák fejlesztésében egyre markánsabb tendenciaként figyelhető meg a nyelvészeti és a kapcsolódó, interdiszciplináris alapkutatások eredményeinek felhasználása a kommunikációelmélet, az információtechnológia, a nyelvtudomány és a pszichológia képviselőinek hatékonyabb eljárások kialakítása érdekében folytatott közös munkájában. Az akusztikai információ ebben az interdiszciplináris perspektívában a mérhető jelekből (alapfrekvencia, intenzitás stb.) származtatott érzetminőségeket (hangmagasság, hangerő, dallam stb.) reprezentáló tulajdonságok olyan együttesévé

² Mint később látni fogjuk, ebben a fonológiai vonatkozásban néhány jellemző (hangszínezet, beszédtempó) háttérbe szorul.

³ „Fontos különbséget mutat azzal a nézettel szemben, amit SELKIRK 1984 és számos további munka képvisel, miszerint a prozódia alapvetően egy reprezentációja a szintaktikai szerkezetnek, azzal együtt, hogy annál laposabb” (Hunyadi, 2003b. 219. o.).

⁴ Ezzel kapcsolatban természetesen fontos azt a szociolingvisztikai tényezőt is figyelembe venni, hogy az élőbeszédre jellemző spontaneitás a modern kommunikáció (közösségi médiák és egyéb virtuális csatornák) világában a *szöveg* médiumát is egyre inkább áthatja.

válik, amely az emberi észlelés (és kogníció) számára is releváns, a kommunikációban funkciót betöltő jelként hasznosul.

A prozódia a beszédtechnológia számos területén (beszélőazonosítás, beszéd- és érzelemfelismerés, a beszéd szegmentálása, beszédszintetizálás) játszik moduláris, néhol kulcsfontosságú szerepet. A gépi tanuláson alapuló rendszerek fejlesztése során több esetben (Shriberg et al., 2000; Szaszák, 2010) nyert igazolást, hogy a prozódiai információk hasznos jellemzőként alkalmazhatók. A vonatkozó kutatások egyik fontos, leginkább távlati és alkalmazásorientált célja az ember–gép kommunikáció természetességének és hatékonyságának növelése. Gördülékenyebben kezelhetővé tehetők például a fordulóváltások, ha a felhasználó közléseinek interpretációjában, illetve válaszok generálásában a beszéd szupraszegmentális sajátosságai is szerephez jutnak (Hirschberg, 2002).

Egy kevésbé feltérképezett, de a fönt vázolt technológiai távlatok szempontjából korántsem érdektelen kutatási területhez sorolhatók azok a prozódia–pragmatika interfész vizsgálatára irányuló munkák (Zellers & Post, 2009; Nakajima & Allen, 1997), amelyek az élőbeszédbeli társas interakciók tematikus és prozódiai struktúrája között igyekeztek korrelációkat kimutatni. Nakajima & Allen (1997) kooperatív dialógusokban⁵ vizsgálták a témaszerkezet lehetséges prozódiai kísérőjegyeit, melynek során összefüggéseket állapítottak meg a megnyilatkozások kezdetén és végén mért alaphfrekvencia-értékek, illetve a környező megnyilatkozásokhoz fűződő propozicionális viszony⁶ között. Az új társalgási témát bevezető megnyilatkozások elején az alaphfrekvencia magasabb és szélesebb terjedelemben változó értékeket mutatott, mint amit a téma folytatásaként később következő megnyilatkozásoknál mértek. Zellers & Post (2009) az alaphfrekvencián kívül egyéb prozódiai jellemzőket is vizsgálat alá vonva megállapították, hogy a tematikus szerkezettel való korrelációk feltárásához a jellemzők együttes vizsgálatára van szükség azok diszkrét elemzése helyett. A feltárt korrelációkból meglátásuk szerint így sem következik biztosan, hogy bizonyos prozódiai minták egyértelmű propozicionális vagy pragmatikai vonatkozással bírnak. Baiat & Szekrényes (2012) egy esettanulmányban megállapította, hogy számukra ismeretlen (magyar, illetve perzsa) nyelven folyó dialógusok hallgatása közben

⁵ A kísérletben részt vevő személyeknek közösen kellett logisztikai feladatokat megoldani egy fiktív város-térkép alapján. A kísérlet ideje alatt külön helyiségben tartózkodtak, fejhallgató és mikrofon segítségével tudtak kommunikálni egymással, amelynek hanganyagát rögzítették a további elemzéshez.

⁶ Az egymást követő megnyilatkozások tartalmi viszonyai szerint különböző kategóriákat állítottak fel: új téma bevezetése, új információ bevezetése a témán belül, előző megnyilatkozás kifejtése, téma folytatása.

a kísérleti alanyok meglepően jó arányban képesek érzékelni a társalgási téma megváltozását pusztán a dialógus dinamikai és prozódiai sajátosságaira hagyatkozva.⁷

1.2. A kutatás célja

A jelen értekezés célja, hogy a fentebb bemutatott előzmények nyomvonalán haladva a beszélt nyelv kutatásához olyan újszerű módszertani eszközökkel és kísérletekkel járuljon hozzá, amelyek a beszéd prozódiai jelenségeit, mint a kommunikáció folyamatának önmagában is informatív akusztikai vetületét a percepció és a konverzációs analízis tükrében vizsgálják. A kutatás keretében megvalósított kísérletek háttérében az alábbi, intuitív hipotézis állt:

A prozódia az élőbeszédben nem csak a megnyilatkozások szintjén szerveződik, nem csak azok lehetséges szemantikai interpretációit modulálja, hanem az interakcióknak egyféle "rugalmas", kvázi zenei közeget biztosítva olyan önálló, bár a verbálishez képest alulspecifikált információtartalommal rendelkezik, ami a percepció számára az interakció értelmezésekor akkor is relevanciával bír, ha a konkrét verbális tartalom valamiért nem hozzáférhető.

Ez a prozódiában kifejeződő, alul-specifikált információtartalom akkor válik igazán érdekessé, ha a társas interakciók tényleges jellemzőinek és folyamatának rekonstruálásához adekvát kiindulópontot is jelent, ami alatt itt természetesen nem a konkrétumokat, a társalgás tárgyalási univerzumának pontos feltérképezését, hanem azokat a különböző intonációs mintákban, dinamikai változásokban, a beszélőváltás ritmusában tetten érhető, a *kommunikációs partnerek viszonyával, szerepével, a beszédhelyzettel* kapcsolatos globális tulajdonságokat, illetve a konverzáció folyamatának lokális eseményeit (pl. témaváltás és egyéb beszédaktusok) értem, amelyek bináris döntések formájában (formális vagy informális, megváltozott a téma vagy nem), a specifikus tartalom ismerete nélkül is kategorizálhatók.

A módszertan oldaláról a disszertáció legfontosabb célját egy egyszerűen algoritmizálható annotációs modell kidolgozása jelentette, ami a kutatás közvetlen kontextusát adó HuComTech projekt (Hunyadi et al., 2012) egyéb elemzési sémái közé integrálva a beszélt nyelv prozódiai jelenségeinek lejegyzésével egészítik ki a korpuszban tárolt interakciók nem-verbális szerkezetét leíró adatokat. A kutatócsoport szempontjából ezek az információk a multimodális kommunikáció természetével kapcsolatos, az interakciókat az egyes modalitások együttműködése és

⁷ Egyéb, szemantikai természetű információkra egy nyilvánvaló akadály miatt nem támaszkodhattak: nem ismerték a nyelvet, amin a dialógus zajlott.

formális alapszerkezete (Hunyadi, 2011) alapján karakterizáló vizsgálatok kiterjesztése miatt voltak érdekesek. Az automatikus prozódiai annotáció kialakításánál fontos szempontot jelentett, hogy kizárólag a hangfelvételek akusztikai elemzésére és a beszédfordulók a korpuszban manuálisan annotált szerkezetére támaszkodjon. Ezt nem csak az az esetleges tényező motíválta, hogy a fejlesztések idején a transzkripciók szótag- vagy tagmondatszintű szegmentálása nem volt teljes mértékben kivitelezett, de a megnyilatkozások prozódiai szerkezetének elemzését elméletsemlegességre törekedve eleve függetleníteni akartam minden olyan szintaktikai és fonológiai szemponttól, amelyek javarészt csak manuális elemzés révén reprodukálhatók, másrészt előfeltételezik a szuprasegmentumok egyéb nyelvi szinteknek való alárendeltségét. További fontos szempontot jelentett, hogy az annotáció kimenete a prozódia fizikailag mérhető változásait perceptuálisan is releváns, diszkrét kategóriákkal felcímkézhető kontúrok formájában reprezentálja. Mivel a fönti elvek alapján kidolgozott annotációs modell a prozódia deskriptív, nem nyelvspecifikus, kontúrelemző feldolgozását célozta meg, ez lehetőséget teremtett arra, hogy az algoritmus implementálása során tanítóanyag elkészítését nem igénylő, szabály-alapú megközelítéseket alkalmazzak.

A értekezés keretében megvalósított percepciók, illetve neuronhálókkal végzett kísérletek célja, hogy a prozódia társalgás-szintű, a verbális tartalomtól független informativitásával kapcsolatos alaphipotézist konkrét klasszifikációs feladatok megoldásával igazolja. A prozódiai annotáció kimenetén tanított neurális hálózatok esetén tehát nem az összes elérhető jellemzőt felhasználó és a maximális hatékonyságot garantáló eljárás kivitelezése, hanem annak a tesztelése jelentette az elsődleges szempontot, hogy a verbális szemantikai tartalom és a vizualitás kizárásával, pusztán prozódia információkra támaszkodva milyen eredmények érhetők el.

1.3. A disszertáció szerkezeti felépítése

A kutatás hátterének bővebb tárgyalása két külön részre tagolódik. A 2. fejezetben kerülnek tárgyalásra a vonatkozó nyelvészeti elméletek, a 3. fejezet pedig a beszédtechnológia a saját kutatásaimban is érintett területeit, illetve a prozódia modellezésére eddig kidolgozott megoldásait járja körbe. A 4. fejezet az disszertáció legfontosabb módszertani eredményét jelentő annotációs eljárást, illetve a fejlesztés elsődleges forrásanyagát jelentő HuComTech korpuszt mutatja be. Az 5. fejezetben gyűjtöttem össze azoknak a percepciók kísérleteknek a leírását, amelyeket az élőbeszéd prozódiai szerkezetére vonatkozó alaphipotézis ellenőrzése céljából bonyolítottam le. A neurális hálózatokkal azonos célból végrehajtott kísérleteket a 6. fejezetben összegeztem.

2. FEJEZET

A kutatás elméleti háttere

Az alábbi fejezetben a beszéd prozódia jellemzőit különféle nyelvészeti elméletek tükrében ismertetem, a téma olyan aspektusait is megszólaltatva, amelyek a konkrét kutatásaim szempontjából talán kevésbé voltak relevánsnak, ugyanakkor kiválóan demonstrálják, hogy a beszéd egy roppant összetett, változatos funkciókkal bíró, egymástól gyökeresen eltérő modellekből is megközelíthető jelenségével állunk szemben. Egy általánosabb bevezetés után a generatív grammatika, majd a beszédaktus-elmélet fényében tekintem át a témát, végül az utolsó alfejezetben („A társalgás prozodiájáról”) a leginkább kapcsolódó kutatások rövid ismertetése mellett tovább részletezem a disszertáció elején megfogalmazott hipotézisem is.

2.1. A prozódiai jellemzők természetéről

A beszéd szupraszegmentális (avagy prozódiai¹) szerkezete és e szerkezetet felépítő jellemzők meghatározása után kutatva a szakirodalomban zavarba ejtő terminológiai sokszínűséget, bizonyos esetekben pedig egymással szembehelyezkedő definíciókat is találhatunk. A szupraszegmentális szerkezet „szinonimájaként” Markó Alexandra korábbi munkájában (Markó, 2005) is bemutatott, a jelenség más és más aspektusát kidomborító terminológiai változatok (*a beszéd zenei elemei, a beszéd mondatfonetikai eszközei, intonáció, prozódia*) és definíciók megosztottságának és bizonytalanságának egyik fő oka a szupraszegmentális szerkezet részét képező jelenségek rendkívüli összetettsége, amely nem csak fonetikai, de funkcionális (fonológiai és pragmatikai) megközelíthetőségükben, illetve a beszéd szegmentális elemeivel való viszonyukban is tetten érhető.

¹ A disszertáció szövegében a „szupraszegmentális szerkezet” és a „prozódia” kifejezéseket egymás szinonimájaként fogom használni. A idegen nyelvű szakirodalom néha a hanglejtéssel összetéveszthető módon az *intonáció* (intonation) terminust is alkalmazza a prozódiai jellemzők összefoglaló nevéként.

A minden elemre egységesen alkalmazható általános meghatározást későbbre halasztva a prozódia részeként számon tartott jelenségek egyszerű felsorolásával valamivel közelebb juthatunk vizsgálatunk tárgyához, amely tehát a beszéd *dallama* (hanglejtés), *hangfekvése*, *átlaghangereje*, *dinamikája* (hangerőváltozásai), *színezete*, *szünetei*, *tempója*, *ritmusa* és *hangsúlyozása* által kifejezett tulajdonságok együtteseként ragadható meg.² A felsorolt jellemzők fényében pedig nagyon is adekvátnak tűnhet a „beszéd zenei kifejező eszközeiként” történő meghatározás, hiszen ugyanezen tulajdonságok (pl. hangsúly, ritmus, dallam, tempó) a zeneművek jellemzésénél is fontos szempontokat jelentenek. A fönti analógiát vitató szerzők (Bolinger, 1972; Gósy, 2000) a zene szigorúbb időszerkezete, tonalitása és művészi megformáltsága oldaláról hangsúlyozzák a beszéd „zeneiségének” a percepció által is visszaigazolható különbözőségét. Természetesen, ha a zene világán belül figyelembe vesszünk bizonyos, az időszerkezeteket lazábban kezelő, kísérleti, atonális zenei műfajokat³, nyitott kérdés marad, hogy ez a megkülönböztetés a megformáltság tekintetében mennyire élesen határolt, vagy mennyiben válik inkább funkcionális, formálisan nem kezelhető esztétikai problémává.

A „szupraszegmentális szerkezet” kifejezéssel fémjelzett meghatározások a beszéd szegmentális elemeivel való viszonyában igyekeznek megragadni a prozódiai jellemzők alaptermészetét. Ezek között találkozhatunk negatív definíciókkal (Nooteboom, 1997), amelyek leginkább azt hangsúlyozzák, hogy a beszéd prozódiai sajátosságai nem vezethetők le a megnyilatkozásokat alkotó szegmentális elemek (beszédhangok) sorozatából, utóbbiaktól független, azokra mintegy opcionális „többletként” ráépülő (Roach, 2009) jelenségről van szó. Szegmentális egységek ugyanazon sorozata tehát a közlés során eltérő prozódiai jellemzőkkel valósulhat meg, előbbiek fonetikai tulajdonságai által nem meghatározott módon. A függetlenség egy másik aspektusát figyelhetjük meg a prozódiai jellemzők hatóköre kapcsán (Cruttenden, 1986), amely általában (hanghatárookra való tekintet nélkül) több szegmentumon ível át, de bizonyos esetekben ugyanazon prozódiai struktúra (pl. egy „emelkedő” dallammenet) alá egyetlen beszédhang (pl. a kérdésként elhangzó: „Ő?”) vagy azok hosszabb sorozata (pl. a kémiai laboratórium felrobbanását közvetlen megelőző: "Ne tedd bele!") is tartozhat. A beszéd e két struktúrája ugyanakkor az artikuláció és a percepció tekintetében egy közös „hangtesten” osztozódik (Gósy, 2004), képzésük és feldolgozásuk párhuzamosan zajlik a beszéd akusztikumának két eltérő vetületében: míg a szegmentális szerkezet elemei a beszédjel akusztikai paramétereinek

² Rögtön hozzá is kell tennünk azonban, hogy a szakirodalomban a prozódiai jellemzők eltérő listáival, elnevezéseivel és rendezésével is találkozhatunk.

³ Az experimentális zene világából idevágó példaként megemlíthetjük Ligeti György *Articulation* című, 1958-as darabját: <https://www.youtube.com/watch?v=PMjuXkrj1Yc>

(amelyek helyett természetesen a „mögöttes” artikulációs gesztusokat is felsorolhatnánk) hangtípusokat karakterizáló időbeli együttállásaként, addig a szuprasegmentumok bizonyos akusztikai paraméterek (alapfrekvencia és intenzitás) és szegmentális elemek szintagmatikus, időbeli viszonyaiként, váltoásaiként írhatók le (Honbolygó, 2011).

A prozódia elemeit tartalmazó korábbi listánkra visszatérve rögtön látható, hogy a fentebbi meghatározás az olyan dinamikus változásokat, idő- és értékviszonyokat kifejező jellemzők esetén fogadható el, mint a beszéddallam (a hangmagasság változása), a dinamika (hangerőváltozás), a szünet (bizonyos időtartamig tartó jelkimaradás), a tempó (szótagok időbeli távolsága), a hangsúly (környezetből kiemelés) és a ritmus (elemek szabályos, időbeli váltakozása). A többi prozódiai jellemző esetében már nem folyamatszerűségről, szegmentális elemek szintagmatikus kapcsolatáról, hanem bizonyos akusztikai paraméterek pillanatnyi/átlagos értékeiről (pl. átlaghangerő, hangmagasság) vagy minőségéről (hangszínezet) beszélünk, amelyek nem játszanak szerepet hangtípusok megkülönböztetésében, de többletinformációként járulnak hozzá a beszéd szuprasegmentális szerkezetéhez. A hangmagasság, a hangerő és az időtartam a legtöbb szerző rendszerében a csoportosítás alappilléreit is jelenti egyben, amelyek változásaként vagy komplexebb szerveződésekként a többi szuprasegmentum levezethető. Így például Alan Cruttenden három alapvető prozódiai jellemzőként a hangmagasságot (*pitch*), a hangerőt (*loudness*) és az időtartamot (*length*) határozza meg (Cruttenden, 1986). A fonológia irányából közelítő kutatók a prozódia elemeit általában azokra a jellemzőkre redukálják, amelyek a szegmentális szerkezet megkülönböztető jegyeihez hasonlóan diszkrét értékekkel jellemezhetők: „Egy prozodikus eszköz akkor diszkrét, ha kategorikusan (»igen-nem«) elkülönülő formákban jelenik meg. Pl. egy dallam vagy emelkedő, vagy nem emelkedő; egy adott helyen van szünet, vagy nincs szünet; [...]” (Varga, 1994. 472 o.). A megszorítás következménye, hogy a rendszerből kiszorulnak azok a szuprasegmentumok, amelyek kategorikusan nem, csak valamilyen relatív, ordinális skála segítségével leírhatók, mint például a beszédtempó, a hangszínezet és a hangerő változásai. Ugyanígy vizsgálaton kívül maradnak például a hanglejtés dallamformáinak kategorikusan nem megragadható tulajdonságai: a hangmagasság ereszkedésének vagy emelkedésének mértéke.

Akusztikai-fonetikai szempontból a beszédjel fizikai paraméterei közül az *alapfrekvencia*, az *intenzitás* és az *időtartam*, amely a szuprasegmentális szerkezet leírásában fontos, egzaktan mérhető tényezőt jelent.⁴ A prozódia fonetikai vizsgálatánál, egy adott jellemző fizikai paraméterei kapcsán viszont figyelembe kell vennünk, hogy az egyes paraméterek mint fizikai

⁴ A prozódiai jellemzők közül e tekintetben kivételt jelent a hangszínezet, amely fonetikailag a legnehezebben megragadható. Vizsgálata jóval több tényező (zöngemínőség, formánsok stb.) bevonását követeli meg.

hangingerek nem egymástól izoláltan, hanem egymás hatását modulálva alakítják ki a percepcióban megjelenő „hangérzetet” (Kassai, 1994). A hangmagasságérzetünkre például nem tisztán a beszédjel alapfrekvenciája, de annak időtartama és intenzitása is hatással van, mint ahogyan a hangosságérzetünket is befolyásolja az alapfrekvencia, a kísérleti fonetika és fonológia egyik nagy kihívását jelentő *hangsúly* még összetettebb jelenségéről (Hunyadi, 2003a; Honbolygó, 2011) egyelőre nem is szólva.

A szupraszegmentális szerkezet elemeit a nyelv rendszerében valamilyen szerepet betöltő eszközökként vizsgálva Varga László az alábbi mondatszintű⁵ funkciókat határozza meg: „Véleményünk szerint az intonációhoz [az intonáció itt tágabb értelemben, a prozódia helyett szerepel - Sz. I.] sorolható prozodikus eszközök mondatfunkciói abban állnak, hogy a hordozó szöveggel kölcsönhatásba lépve abból megnyilatkozást (kiejtett mondatot) hoznak létre, miközben tükrözik a mondat attitűdtartalmát, logikai jelentését, tematikus jelentését és kommunikatív mondat típusát.” (Varga, 1994) Az idézetben szereplő „tükrözést” mint bizonyos tartalmak megkülönböztetésére és tagolására irányuló funkciót értelmezve láthatjuk, hogy a prozódia nem csupán az időben egyedi és megismételhetetlen fizikai jelenséggé megvalósuló beszéd szükségszerű velejárója, amely során időtartammal, frekvenciával és intenzitással jellemezhető ingereket állítunk elő, de egyúttal ezek a paraméterek intencionált nyelvi funkcióval bírva a megnyilatkozások értelmezésének is fontos tényezői lesznek mind fonológiai, szintaktikai, logikai és pragmatikai tekintetben.

2.2. A prozódia helye a nyelv és a kommunikáció modelljeiben

2.2.1. „Színtelen zöld ideák intonálnak-e?”

Ha a prozodiát a nyelvtudomány két tágabb, dichotómiákat alkalmazó modelljeiben kívánjuk elhelyezni, akkor a beszéd megnyilatkozásaihoz kötődő természetükből kiindulva Saussure (nyelv és beszéd kettősségén alapuló) rendszerében rövid úton a *parole* (a beszéd), a chomskyánus nyelvfelfogás tekintetében pedig a *performancián* keresztül megnyilvánuló „tágabb értelemben vett nyelvi képesség” (Hauser et al., 2002) hatókörében találjuk meg a megfelelő helyet.⁶

⁵ Ebből a hatóköri meghatározásból eredeztethető a prozódiai jellemzők „mondatfonetikai eszközökként” való elnevezése is, amelynek több szerző (Kugler, Gósy, Markó) problematikusnak talál abból a szempontból, hogy a beszélt nyelvben nem mondatokkal (mint absztrakt nyelvi egységekkel) hanem megnyilatkozásokkal (az előbbieket megvalósulásaival) találkozunk.

⁶ Ezzel a párhuzamos megközelítéssel nem állítjuk, hogy Saussure és Chomsky rendszere tökéletesen megfeleltethető volna egymásnak. Párhuzamba állításuk alapját itt kizárólag a prozódia státuszának kérdése jeleneti. Az „azonos nézetek különbségeivel” kapcsolatban lásd Németh T. Enikő kapcsolódó tanulmányát (Németh, 2006).

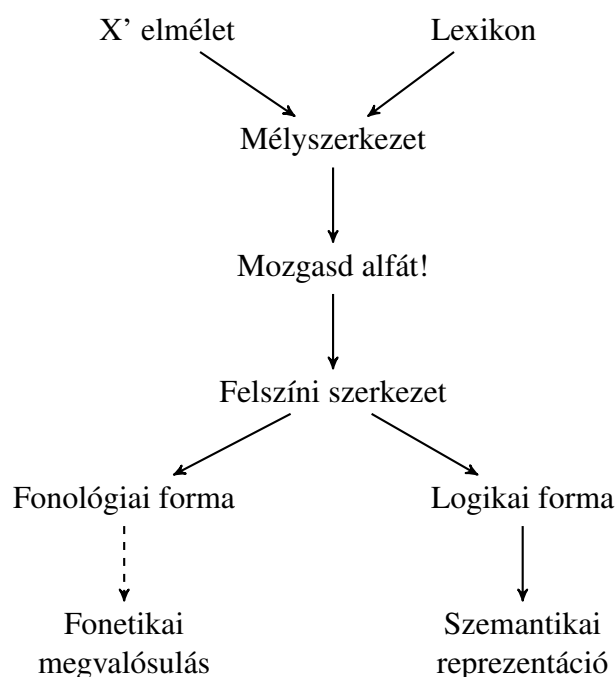
A nyelvi szerkezeteket megvalósító megnyilatkozások prozódiajának elemei (a kiejtett mondat hanglejtése, színezete stb.) ebben a formális megközelítésben tehát nem a nyelv rendszerének (a beszélő nyelvi kompetenciájának) részei, hanem a beszéd sajátosságai (a nyelvi jel jelölő oldalának attribútumai), amelyek nem a *nyelvben*, hanem a *kommunikációban* rendelkeznek disztinktív vagy egyéb funkcióval.⁷ A nyelvi kód ebben a szűkebb (Chomsky esetében szintaxis-központú) értelmezésben az adott grammatika által generálható absztrakt szerkezetek (jól formált mondatok) halmazára korlátozódik,⁸ függetlenül attól, hogy ezek a kognitív produktumok egy adott kommunikációs helyzetben hogyan és milyen kommunikációs céllal valósulnak meg, hiszen egy nyelv lehetséges mondataként (grammatikalitási ítéletünk tárgyaként) fogadhatjuk el jelek egy szekvenciáját akkor is, ha közben nem társítunk hozzá semmilyen jelentést, fizikai paraméterekkel jellemezhető hangalakot vagy kommunikatív funkciót, mivel nem közlésként, hanem egy nyelv absztrakt elemeként (mondataként) hivatkozunk rá. Chomsky elhíresült, értelmes jelentéssel nem rendelkező, de teljesen grammatikus példamondata („Színtelen, zöld ideák dühödten alszanak.”) jól demonstrálja ezt,⁹ de vegyünk most helyette egy másik, a prozodiát jobban játékba hozó példát: "Milyen színű ez a madár (?)". Szavaknak ugyanezt a szekvenciáját különböző helyzetekben, eltérő hanglejtéssel is használhatjuk úgy, hogy közben a mögöttes nyelvi struktúra változatlan marad, azonban az eltérő hanglejtés, eltérő kommunikációs szándékra utal. Egyik esetben valóban a madár különös színének nevére vagyunk kíváncsiak, egy másik szituációban pedig, tegyük fel, egy rajzoló gyermek tévedésére, a választott és a madár valószínű színe közötti eltérésre próbáljuk ráirányítani a figyelmet. A nyelv és a kommunikáció gyakori együttjárása ellenére is független volta ugyanakkor visszafelé is igaz: a prozódia kommunikációs funkciókat tölthet be nyelvi tartalom (grammatikailag értelmezhető nyelvi kód) hiányában is. Például egy emelkedő dallammenettel kiejtett artikulálatlan hangsor (pl. "Hmmmmmm?") adott helyzetben jelentéssel bíró közlésként értelmezhető.

A nyelv generatív modelljének fonológiai komponense – amelyben a szupraszegmentumok, mint prozodémák szerephez juthatnak – egy a szintaxisnak alárendelt, a felszíni szerkezet fonológiai formáját a szemantikai reprezentáció alapját jelentő logikai formától függetlenül előállító modulként kap helyet. Mint a 2.1. ábrán is látható, ebben a modellben mind a két komponens

⁷ A főnti megállapítás Chomsky rendszerében nyelv és kommunikáció nemazonosságának azon tézisének alapul, miszerint a nyelvi szerkezetek elsődleges funkciója nem a *közlés* (mint kommunikációs funkció), hanem a *gondolat* megformálása (mint kognitív funkció).

⁸ Amellett, hogy a lehetséges szerkezetek száma az újraíró szabályok rekurzív alkalmazhatóságának köszönhetően potenciálisan végtelen.

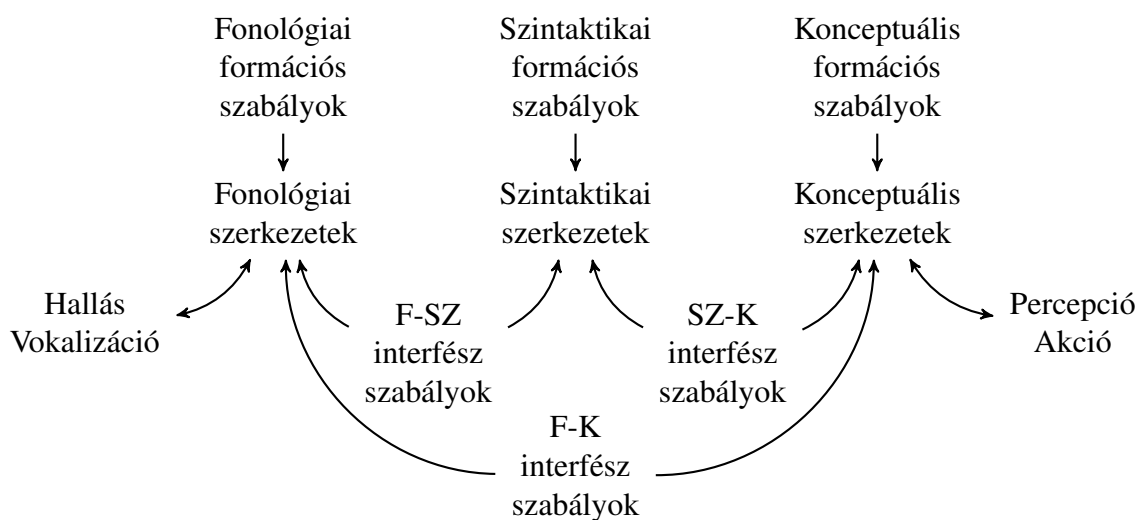
⁹ A nyelv logikai pozitívista szemléletének követői az ilyen szerkezetek lehetőségességét, a természetes nyelvi grammatika szemantikai „vakságát” tették felelőssé számos, metafizikainak bélyegzett látszatállítást megszületésében. Ezzel kapcsolatban lásd Rudolf Carnap tanulmányát (Carnap, 1972).



2.1. ábra. A nyelv generatív architektúrája (Chomsky 1981).

(fonológia és szemantika) ugyanazt a szintaxis generálta szerkezetet interpretálja tovább a saját szabályai szerint, anélkül, hogy egymás kimeneteire hatással lennének. A korábbi modellekben a szemantikai reprezentáció fonológiai formától való függetlenségét még inkább hangsúlyozta, hogy az előbbi közvetlenül a mélyszerkezethez társult hozzá. A szintaxisban tehát úgymond minden eldől: a nyelvi forma jelentése és a későbbi minimalista programban a „spell out” hatókörébe kerülő fonológiai formája is. Mindez a generatív grammatika azon törekvésével indokolható, hogy modelljével nem a kommunikáció folyamatát igyekszik megragadni, melyben bizonyos intonációval kiejtett megnyilatkozásokhoz társítunk jelentéseket, hanem a nyelv működését egy lehetséges szerkezetek generálására (és felismerésére) képes eszközként, kognitív képességként írja le. A paradigma célja tehát nem a nyelvi szerkezetek konkrét kommunikációs helyzetekben történő alkalmazásának (a performanciának) bemutatása, hanem a lehetséges szerkezeteket körülhatárolása olyan szabályok megalkotásával, amelyek segítségével ezek a szerkezetek generálhatók, amit a modellben egyedül a szintaxis újraíró szabályainak rekurzív alkalmazhatósága képes biztosítani. A „szűkebb értelemben vett nyelvi képesség”, a nyelvi kreativitás kognitív motorja tehát a lehetséges szintaktikai szerkezetek potenciális végtelenségét¹⁰ implikáló rekurzió lesz, amelynek az esetleges fonetikai megvalósítás csak a performanciában szab korlátokat.

¹⁰ A végtelenség itt kettős értelemben szerepel: végtelen számú és elméletben (pl. a negáció ismételt alkalmazásával) végtelen (korlátlan) hosszú szerkezet generálható.



2.2. ábra. Jackendoff párhuzamos architektúrái (Jackendoff, 2002).

A generatív grammatika szintaxis-központúságát igyekszik oldani Ray Jackendoff párhuzamos architektúrákban gondolkodó modellje (Jackendoff, 2002), amelyben a fonológia, a szintaxis és a szemantika önálló szerkezetek generálására alkalmas, egymásra kölcsönös hatással lévő modulként kap helyet (lásd 2.2. ábra). Az egyes modulok formációs szabályai ebben a modellben a szintaxisnak nem alárendelve, hanem önállóan határozzák meg a lehetséges fonológiai, szintaktikai és konceptuális szerkezetek halmazát, és ami ennél is fontosabb, hogy habár bizonyos (a 2.2. ábrán „F-SZ” és „SZ-K” rövidítésekkel jelölt) interfész szabályokon keresztül a szintaktikai forma itt is szabályozni (korlátozni) fogja a felszíni szerkezethez társítható fonológiai és szemantikai struktúrákat, ez a szabályozás kölcsönös (az interfész szabályok a szintaktikai szerkezetekre is hatással lesznek), illetve a hagyományos generatív modellből hiányzó fonológia-szemantika (a 2.2. ábrán „F-K” rövidítéssel jelölt) interfészen keresztül lehetővé válik az eddig szintaxisnak alárendelt modulok, a fonológiai és a konceptuális szerkezet egymásra hatása is. A két komponens közvetlen kapcsolatának hangsúlyozását többek között azok a performanciától független nyelvi jelenségek indokolják, amikor két egymással oppozícióba helyezhető konceptuális szerkezet különbsége (pl. eltérő elemek kerülnek fókusz és topik pozícióba) a szintaxis eszközeivel nem, csak a fonológiai forma által (pl. hangsúlyozás segítségével) kifejezhető (Hunyadi, 2003b). Továbbá technikailag elegánsabb és feltehetően plaubizilisebb kognitív magyarázatot kapunk az egyes mondatok grammatikalitására is, ha különböző modulokat mint a párhuzamos feldolgozás moduljait tételezzük fel (Hunyadi et al., 2014).

2.2.2. Megnyilatkozások és beszédaktusok

A nyelvi struktúrák absztrakt dimenziója felől a performancia, a nyelvi jelek kommunikációban történő használata és megvalósulása felé közelítve a *mondatok*, habár továbbra sem szűnnek meg egy grammatika produktumaiként egzisztálni, *megnyilatkozásként* vizsgálva őket jelentésük egyéb kommunikatív funkciókkal, a jelek formális és konvencionális jelentésén túlmutató pragmatikai olvasattal gazdagodva a beszédaktus-elmélet (Austin, 1962; Searle, 1969) keretein belül ragadhatók meg. Az elmélet gyökerei egészen John L. Austin munkásságáig nyúlnak vissza, aki a megnyilatkozások világát funkcionális szempontból az ún. konstatív és performatív aktusokra osztotta fel, ahol a performatívumok a klasszikus logikai elemzés tárgyát képező kijelentések, a leírásra és a tényközlésre szorítkozó konstatívumok információtartalma mellett cselekvésértékkel is rendelkeznek. Ez a felosztás a későbbiekben elvesztette relevanciáját, hiszen az információközlés, még ha a performatív aktusok prototipikusabb eseteinek tekinthető direktívumokhoz képest kevésbé explicit formában is, de értelmezhető úgy, mint egy cselekvés végrehajtása ('*x azt mondta y-nak, hogy z*').

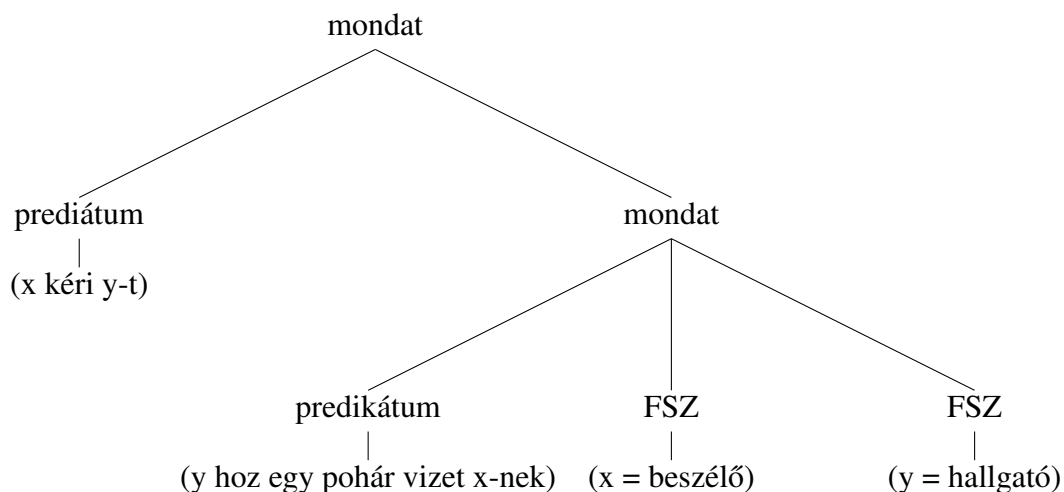
A beszédaktusok két fő komponensre oszthatók fel: a lokúciós aktus (a kimondott hangsor) logikai értelemben vett proposícionális tartalmára és az ehhez (afféle funktorként) hozzátársuló, az aktus cselekvésértékét és a beszélő szándékát kifejező illokúciós aktusra. Egy egyszerű példával élve a „Holnap segíteni fogok kitakarítani” megnyilatkozás proposícionális tartalma az alábbi, a jelenben nem verifikálható jövőbeli tény: '*[a holnapi nap] x segít kitakarítani y-nak*'¹¹, ami a megnyilatkozás formális, konvencionális jelentését is kifejezi, de amihez még hozzátársul az ígérettevés, az elköteleződés¹² illokúciós aktusa is. A kijelentés adekvát értelmezéséhez a megfelelő kontextus (y épp takarít) és *intonáció* esetén hozzátartozik az a lehetséges, indirekt közlésként való interpretáció is, amelyben az „elköteleződés” csak másodlagos aktus, az elsődleges illokúció pedig egy rejtett, az alábbi módon explikálható direktívum: „Hagyd abba a takarítást! Holnap majd megcsináljuk együtt!”. Ez utóbbi esetben a megnyilatkozást sikeresnek¹³ tekinthetjük, ha az érintett hallgató fél az aktus őszinteségi feltételének meglétében bízva felfüggeszti a környezet tisztasága érdekében folytatott tevékenységét¹⁴.

¹¹ A változók, így a „holnapi nap” a kontextusnak megfelelően helyettesítendőek.

¹² Searle rendszerében ezeket a beszédaktusokat *komiszívumoknak* hívjuk, amelyek a világot próbálják az ígéretben megfogalmazott szavakhoz igazítani. Igazságértékkel nem rendelkeznek, csak megvalósítottak, betartottak, vagy megszegettek lehetnek.

¹³ A performatívumok esetében logikai igazságértékek helyett sikerességről vagy sikertelenségről beszélhetünk, ami különféle feltételek (adekvátsági, őszinteségi stb.) teljesülésének vagy nem-teljesülésének függvénye.

¹⁴ A közlés hallgatóra gyakorolt tényleges hatását a beszédaktus-elméletben „perlokúciós aktusnak” nevezik.



2.3. ábra. A „Hozzál nekem egy pohár vizet!” direktívum mélyszerkezete (Kiefer, 2009).

Kevésbé indirekt közlések esetén megközelítés kérdése, hogy a formális szemantika elsődleges tárgyát képező, kizárólag a nyelvi összetevők (szótár és szintaxis) által meghatározott, *mondatokhoz* társítható és a beszédaktus-elmélet keretében vizsgált pragmatikai jelentést egymással rivalizáló paradigmák produktumainak tekintjük, vagy Kiefer Ferenc tanulmánya (Kiefer, 2009) nyomán a „nyelvészeti pragmatikát” az eszközként használt nyelvi struktúrák mélyszerkezetének feltárásához való konstruktív hozzájárulásként építjük be a formális modellekbe. Kiefer példájában a „Hozzál nekem egy pohár vizet!” direktívum mondatának mélyszerkezete az illokúciós aktust, mint predikátumot tartalmazva teszi explicitté az aktusban rejlő logikai viszonyokat (lásd 2.3. ábra). Általánosabb, kvantifikált formában megfogalmazva egy beszédaktus esetén a felszíni szerkezetben kifejezett propozicionális tartalom (P) mindig valamilyen illokúciós aktus (I_{act} , például: „ x azt kéri y -tól, hogy...”.) bemenete lesz, többnyire azonos névparamétereket váró predikátumként ágyazódva be:

$$\exists x \exists y : I_{act}(x)(y)(P(x)(y)) \quad (2.1)$$

Főként indirekt beszédaktusok¹⁵ esetében azonban a fentiekhez hasonló, pragmatikai igényű elemzések nem az eszközként használt nyelvi szerkezet, hanem a beszélő szándékát kifejező preverbális üzenet (kommunikálandó kognitív tartalom) jelentését kívánják megragadni, amiben – az illokúciós aktus által kifejezett szándék megfelelő értelmezésében – a megnyilatkozás prozódíája (a disszertáció szempontjából különösen) fontos támpontot jelent. További különbség, hogy egy beszédaktus „jól formáltságában” (más szóval: *adekváttságában* vagy si-

¹⁵ A fentebbi, ígéretnek „öltöztetett”, takarításra vonatkozó *kérés* mellett tankönyvi példának tekinthető az „Ide tudnád adni a sőt?” fordulat, amelynek adekvát interpretációja egy konkrét beszédhelyzetben természetesen nem a hallgató átnyújtásbeli képességeit firtató kérdésként vezethető le.

kerében) a felszíni szerkezetben közölt proposícionális tartalom logikai igazságértéke, illetve a használt nyelvi szerkezetnek a grammatikalitása helyett a pragmatikában egészen más szempontok kerülnek előtérbe. Mivel az esetek többségében a proposícionális tartalom egy intenzionális funktor argumentumában szerepel¹⁶, a bemenet igazságértéke a logika törvényei szerint sem határozza meg egyértelműen a kimenet (a beszédaktus) „igazságértékét”, illetve a beszélő megnyilatkozásának agrammatikussága a hallgatót nem feltétlen akadályozza meg az eredeti üzenet visszafejtésében. Ugyanakkor a beszédhelyzetben elvárható vagy az egyértelműsítéshez szükséges prosódiai megformálás hiánya könnyen az üzenet interpretációjának sérülését vagy a társalgás megakadását okozhatja („Ezt most kérdezed vagy mondod?”).

A beszédprodukció Levelt által kidolgozott általános modelljében (Levelt, 1989) a megnyilatkozások értelmezésére (avagy percepciójára) vonatkozó megállapításainkhoz hasonló szervezési elvekkel találkozhatunk. Az úgynevezett *prosódia generátor* egyik bemeneti forrása a *felszíni szerkezet*, amely a fogalmi tervezés szintjén létrejött *preverbális* üzenet (konceptuális szerkezet) grammatikailag átkódolt (a generatív grammatika azonos elnevezésű szintjével rokonítható) reprezentációja. Szintén fontos bemeneti forrást jelent a fonológiai (és morfológiai) kódolás egymást követő szintjein megszülető *metrikai* és *szegmentális* szerkezet, melyekhez utolsó, afféle retorikai összetevőként a beszélő szándékait (a közlésben rejlő illokúciós aktust) és érzelmeit kifejező *intonációs jelentés* társul hozzá. A prosódia tehát nem csak a beszéd percepciójában fontos tényező, de az értelmezés folyamatának megkönnyítése célján a beszéd produkciójának is szándékolt eleme, ha a megformálás tekintetében ez nem is mindig zajlik úgy-mond „tudatosan”.

2.3. A társalgás prosodiájáról

Mint az értekezés bevezetőjében ez már említésre került, a disszertáció keretében megvalósított kísérletek háttérében az a hipotézis állt, miszerint prosódiai információkra nem csak egy megnyilatkozás nyelvi formájának adekvát interpretációjában támaszkodhatunk, de a résztvevők között zajló interakció egészének az elemzésében is releváns kiindulópontot jelenthet akár azokban az esetekben is, amikor a verbális tartalom valamilyen oknál fogva nem hozzáférhető. Ebben a megközelítésben viszont a prosodiát már nem egy grammatikai rendszerben betöltött

¹⁶ Ilyen intenzionális kontextusnak számít például a felszólítás aktusa is („Kérem, mindenki álljon fel fel!”), ahol a felszólítás „igazságának” (adekvátságának) nem előfeltétele a proposícionális tartalom („Mindenkinek áll”) igazsága. Az intenzionalitás szempontjából ugyanakkor érdekes, hogy az adekvátság pontosan akkor válik kérdésessé, ha a proposícionális tartalom igaznak bizonyul (Miért kérném, hogy álljanak föl, ha már mindenki áll?).

(fonológiai) funkciói, hanem a beszédaktusok megvalósításának (a beszélő intencióit kifejezésre juttató) eszközöként a pragmatika, illetve a nem intencionált tartalmak (pl. a beszélő érzelmi állapota, egyéni beszédmódja stb.) vonatkozásában a paralingvisztika irányából vizsgáljuk. Nem azok az esetek lesznek tehát érdekesek, amikor bizonyos, a szemantikai szerkezetet érintő különbségek (pl. topik és a fókusz pozíciói, a kvantorok hatóköre) a felszíni szerkezetben a prozódia segítségével válnak explicitté (Hunyadi, 2003a), hanem amikor a prozódia a nyelvi formától (a verbális tartalomtól) függetlenül közvetít a konkrét szemantikai tartalomhoz képest alul-specifikált, de az interakció gördülékeny lefolyásához és értelmezéséhez szükséges információkat.

A téma szakirodalma leginkább a fordulóváltások prozódiai sajátosságaival, a társalgások temporális szerveződésével foglalkozik, amelynek elméleti alapjai és fogalomrendszere még a hetvenes években, Sacks, Schegloff és Jefferson (Sacks et al., 1974) nyomán születtek meg, de a szóátadás jelzésére vonatkozó elméletek és a prozódia szerepének hangsúlyozása kapcsán fontos kiemelniünk Duncan (1972) munkáját is. A fordulóváltásokkal kapcsolatban sokat foglalkoztató kérdés, hogy az interakció résztvevői a társalgás során milyen szempontok alapján választják ki az úgynevezett „beszélőváltásra megfelelő helyeket” (Iványi, 2001). Az aktuális forduló birtokosának nem feltétlen szükséges direkt verbális gesztusokkal explicitté tenni („Várj még ne szólj közbe, mindjárt folytatom”), hogy az addig közölteket még további kommentekkel szándékozik kiegészíteni, vagy ellenkező esetben felkínálni kommunikációs partnerüknek a fordulóátvétel lehetőségét („Befejeztem, átveheted a szót”). A forduló befejezésére vagy folytatására a megnyilatkozás megformálására használt intonáció alapján is tehetünk predikciókat (amelyek sikere természetesen nem jelenti azt, hogy a beszélgető partner ehhez minden körülmények között igazodni is fog). A forduló megtartására vonatkozó beszélői szándék kifejeződhet például a Varga rendszerében (Varga, 1994) „féleső”, „eső-emelkedő” (a funkció szempontjából önálló, de nem befejezett), vagy akár a „lebegő jellegű” (a funkció szempontjából előremutató) címkékkel jellemzett karakterdallamokkal. Ez utóbbival rokonítható a ToBI (Silverman et al., 1992) lejegyzési modelljében szereplő L-H% intonációs frázis, amely az úgynevezett lista elemek („LIST ITEM”) megformálására szolgál. Egy kiskereskedelmi kontextusban zajló vásárlási szituációban például mindaddig ezt formát használom, amíg a választott termékek listájának végére nem érek, ezáltal jelezve az eladónak, hogy még nem időszzerű a felsorolás lezártságának megerősítését kérő „Lesz ezen kívül még valami más?” megnyilatkozással félbeszakítania. A kérdéses kontextusban érdekes módon az intonáció sokkal megbízhatóbb alapul szolgál a lezárásra vonatkozó predikcióhoz, mint például a felsorolás elemei között tartott szünetek hossza, mivel az utóbbi nagyobb valószínűséggel utal a gon-

dolkodási idő kérésére. Az intonáción kívül azonban más, nem csak az aktuális fordulóban megfigyelhető prosódiai jellemzők segíthetik a társalgás gördülékenységét, a fordulók értelmezését.¹⁷ Az egyes megnyilatkozások lokális jellemzőin túl, magának a fordulóváltásnak is van egy úgynevezett ritmusa (Couper-Kuhlen, 1991), amely a prosódiát egy tágabb perspektívában megvilágítva (Couper-Kuhlen & Selting, 1996) sokat elárulhat a az interakciónak keretet adó szituációról és a résztvevők viszonyáról is.

A fordulóváltások mikrostruktúráján túl egy másik fontos kutatási témát jelent a konverzáció makrostruktúrájának, globális egységeinek az elemzése (Clarke & Argyle, 1997; Ping, 2005). A prosódia ide vonatkozó, a társalgás tematikus tagolásában betöltött szerepe kevésbé tűnik evidensnek, de mint a bevezetőben már említettük, találhatunk olyan kutatásokat is (Nakajima & Allen, 1997; Zellers & Post, 2009; Zellers, 2012), amelyek a jelen értékezéshez hasonlóan ezzel a kérdéssel foglalkoznak. Hazai viszonylatban pedig egy egészen friss fejlemény, hogy a BEA beszédkorpusz (Gósy et al., 2012) alapján szintén történtek olyan pilot vizsgálatok (Hámori & Horváth, 2019), amelyek nem csak a sokat kutatott (Markó & Gósy, 2015) beszédfordulók szintjén elemezték az interakciók prosódia szerkezetét, hanem a három résztvevővel folyó társalgásokat nagyobb tematikus blokkokra tagolták. A mi kérdésünk most tehát az, hogy tehetünk-e prosódia alapján a fordulóváltáshoz hasonló predikciókat egy társas interakció témájának megváltozására is? Más szóval egy diskurzus fordulóváltozásai közül a prosódiai szerkezet alapján kiemelhetők-e azok a speciális esetek, amelyek egyúttal a társalgási téma megváltozását is maguk után vonják (különböző témákra szegmentálják azt), így lehetővé téve az interakció tematikus szerkezetének a konkrét verbális tartalomtól független rekonstrukcióját? Első körben a „témaváltás” pontos jelentését kell, hogy meghatározzuk, ami egy formális megközelítésben felfogható a beszélgetés tárgyalási univerzumának egyfajta felfrissüléseként, ami új diskurzus referenseket bevezetve megváltoztatja bizonyos kifejezések (pl. névmások) referenciáját, illetve kijelöli a megváltozott kontextus szempontjából adekvát megnyilatkozások halmazát.¹⁸ Természetesen nem ezek az absztrakt, „zajtalanul” történő szemantikai változások azok, amik a környező megnyilatkozások prosódiai szerkezetében nyomot hagyhatnak. Azonban – hipotézisem szerint – a változást kiváltó, a társalgásnak új irányt kijelölő kommunikációs gesztus, és a többi partnertől erre érkező reakció már hatással lehet az interakció „zenei atmoszférájára” is. Ez főként azokban az esetekben valószínűsíthető, amikor

¹⁷ A beszédtempó ide vonatkozó lehetséges szerepeivel kapcsolatban lásd Plug (2009).

¹⁸ Ha például egy személyes témákat boncolgató diskurzus után megkérem kollégámat, hogy térjünk vissza az aktuális teendők megvitatására, inadekvát reakcióként konstatálnék egy, a szüleim hogylétére irányuló kérdést.

a témaváltás az interakció természetes menete által nem motivált módon, hanem váratlanul, vagy mint az interjúszituációk esetén, egy előre eltervezett forgatókönyvnek megfelelően következik be, ami a tárgyalási univerzum drasztikus, az előző állapothoz képest radikális eltérést eredményező felfrissülését vonja maga után. Egy hétköznapi beszédhelyzetben hasonló mértékű átszerveződéssel járhat az aktuális téma kimerülése vagy a további kifejtés problematikussá válása (pl. egy új, a tárgyalási univerzumot még nem ismerő résztvevő becsatlakozása miatt). A témaváltás „motivátlanságát” tehát szerencsésebb úgy értelmezni, hogy ez a motiváció nem az előző téma tartalmából, hanem egyéb külső szempontokból táplálkozik. Feltételezésem szerint akármelyik esetről is legyen szó, az új témára való áttérés, illetve valamelyik fél által történő átvezetés egy olyan „hatalmi” aktusként tartható számon, ami ideális esetben igyekszik a megfelelő időzítés és forma adekvátsági kritériumának is megfelelni, ami (főként a motivátlatlan esetekben) nem feltétlen szemantikai, hanem az átvezetés aktusát a társalgás addigi sodrásához, zenei szerveződéséhez igazítva, illetve az új irány tényére a hanghordozással öntudatlanul is reflektálva prozódiai szempontokat alkalmaz.

Ha a beszédprodukció oldalán feltételezzük is ezeknek a prozódiai szempontoknak a relevanciáját, továbbra is kérdés marad (és a disszertáció keretében lebonyolított kísérletek tulajdonképpen erre keresik a választ), hogy egy olyan speciális megfigyelési szituációban, ahol a verbális tartalom nem hozzáférhető (pl. egy ismeretlen nyelven zajló vagy zárt ajtók mögül kiszűrődő beszélgetés) a percepció számára mennyire jelent biztos támpontot a témaváltásra (vagy mondjuk a teljes konverzáció lezárására) irányuló kommunikációs stratégiák prozódiai vetülete ezeknek a eseményeknek a felismerésében? A megfigyelő felismerésménye – függetlenül attól, hogy a tényleges történéseknek megfelel-e – ezekben a speciális esetekben nem az interakcióban résztvevők által szándékolt módon alakul ki. Az életszerű megvilágítása kedvéért vegyük példának *A*, *B* és *C* személyek esetét, ahol *A* és *B* egy zárt helyiségben párbeszédet folytat. A párbeszéd témája *C* bizonyos, a beszélgetés során negatív megvilágításba kerülő múltbéli cselekvései. A téma kellős közepén, hallható módon belép a szomszédos helyiségbe *C*. *A* és *B* – *C* közelségétől tartva – hirtelen (nem a társalgás természetes menetének megfelelő módon) témát vált. *C*, aki ugyan a beszélgetés szavait nem hallja (a szemantikai tartalom közvetlenül nem hozzáférhető), mindössze annak a beszédfoszlányokból is kirajzolódó dinamikájáról rendelkezik képzetekkel, *A* és *B* szándékaival ellentétben (ha ok okozati összefüggésbe nem is feltétlenül hozza belépésével) konstatálja az interakció dinamikájának megtörését, amiből az interakció tematikus folytonosságának sérülésére következtet. A példa természetesen több szempontból is problematikus. *C* benyomásai az interakció dinamikáján túl például nagy mértékben támaszkodhatnak a szituációval, illetve *A* és *B* személyekkel kapcsolatos háttérismereteire is, de hasonlóan

érvelhetünk akár a fordulóváltással kapcsolatban is, ahol az „időkérés” sikerességében a prozódian kívül egyéb vizuális és kontextuális információk is szerepet játszhatnak.¹⁹ Továbbá nyitott kérdésként marad hátra az is, hogy a megfigyelő mennyire képes a benyomások potenciális alapját jelentő prozódiai szerkezetek tulajdonságaira reflektálni. Ha megkérdeznénk az előző példa 'C'-vel jelölt szereplőjét, hogy mi alapján érezte úgy, hogy 'A' és 'B' beszélgetésének témája megváltozott (anélkül, hogy értette volna az elhangzott szavakat), vajon tudna-e érdemben válaszolni, általa megfigyelt jelenségekre hivatkozni? Talán egy, a beszélgetésben bekövetkező hosszabb szünet és az előtte-utána megfigyelhető dinamikai viszonyok speciális aránya segítséget jelentene a megfigyelés verbális alátámasztásában és tudatosításában, de könnyen lehet, hogy hasonló zavarban érezné magát, mintha egy számára ismert személy hangjának felismerésményéről vagy fiziológiai értelemben vett egyensúlyozási képességének alkalmazási módjairól kellene tudatosan számot adnia.

Saját kísérleteim kidolgozása során az elérhető eredményekre és a felhasználható módszerekre nézve negatív hipotézisből indultam ki, miszerint a társas interakciók tematikus szerkezetének mélyebb feltáráshoz nem, vagy csak nagyon korlátozott esetekben²⁰ elegendő a prozódia pusztán megnyilatkozásszintű elemzése és kiértékelése. Például a társalgásban új témát bevezető megnyilatkozások prozódiai sajátosságainak összegyűjtése²¹ rendkívül informatív lehet, a téma változását nem implikáló megnyilatkozások hasonló tulajdonságaival összevetve azonban nem elvárható, hogy bármelyik jellemző (kezdeti alaphangfrekvencia-érték, dallamforma, tempó stb.) vagy akár ezek bizonyos kombinációja olyan – a kísérletek független változóit jelentő – pragmatikai tényezőkre visszavezethető, szignifikáns különbséget mutat, amely alapján a megnyilatkozások tematikus funkciójának („sérti-e” a téma folytonosságát?) kategorizálása biztonsággal elvégezhető. Hogy a kísérletek során megfigyelt benyomások prozódiai alapjáról közelebbi képet kapjunk, hipotézisem szerint célszerű, ha nem csak a megnyilatkozások diszkrét egységeit, hanem a prozódiai szerkezet globálisabb viszonyait és dinamikáját, a megnyilatkozások időbeli határaitól független, a szünetek és a beszélőváltás időszerkezetével is kiegészített *folyamatát* vizsgáljuk.²² A témaváltásra vagy egyéb kommunikációs események pozíciójára

¹⁹ Az általam végzett percepciós kísérletekben ezért kellett gondoskodnom a verbális és a vizuális modalitás kizárásáról, illetve arról, hogy az alanyok ne rendelkezzenek háttérismeretekkel a megfigyelt interakcióra vonatkozóan.

²⁰ Egy megnyilatkozás modalitásának, illokúciós erejének megállapításához néha a hanglejtés önmagában elégséges információt szolgáltathat.

²¹ Nakajima & Allen (1997) kutatásában ezt a megnyilatkozások elején és végén mért alaphangfrekvencia-értékekkel kapcsolatban láttuk.

²² A hipotézist a percepciós kísérletekre is kiterjesztve: az interakciókat megfigyelő személy benyomásai sem csak a társalgás diszkrét, megnyilatkozásszintű egységein, hanem azok folyamatán, előzményein és viszonyain alapulnak.

utaló szerkezeti minták pragmatikai értelemben vett többértelműségét természetesen ez az alternatív módszertani megközelítés sem fogja feloldani. Ilyen többértelműségről beszélhetünk például egy hosszabb szünet és a beszédtempó vagy az intonáció – a szünetet megelőző állapothoz viszonyított – megváltozása esetén, amely alapján talán helyesen következtethetünk témaváltásra, de a jelenséget, a társalgás addig gördülékeny dinamikájának megakadását, prosódiai színezetének megváltozását okozhatja valamilyen egyéb kommunikációs vagy kognitív tényező is, ami nem jár együtt témaváltással. A másik irányból közelítve meg a problémát: a témaváltás bizonyos esetekben megtörténhet anélkül is, hogy bármi észrevehető nyomot hagyna az interakció prosódiai szerkezetében.²³ Így a témaváltás (*TV*) egy bizonyos prosódiai minta (*PM*) megfigyelhetőségén alapuló valószínűségének meghatározásakor a Bayes tétel alapján figyelembe kell vennünk azoknak az eseteknek a valószínűségét is, amikor egy hasonló prosódiai minta témaváltás nélkül fordul elő:

$$P(TV|PM) = \frac{P(PM|TV)P(TV)}{P(PM)} \quad (2.2)$$

A prosódia többértelműségének és az utóbb említett esetekben tapasztalható semlegességének (avagy a szemantika „zajtalanságának”) oka, hogy egy interakció pragmatikai és prosódiai szerkezete közötti összefüggés nem olyan közvetlen és egyértelmű megfeleltetés, mint ami a fonológiában a fonémák és az őket megkülönböztető fonetikai jegyek között meghatározható. A témaváltás esetében például nem a társalgás tematikus koherenciájának újraszerveződése – mint szemantikai jelenség – az, aminek fonetikailag leírható jegyei vannak, hanem a témaváltáshoz opcionálisan, a szituáció függvényében alkalmazott társalgási stratégiák, a korábbi téma esetleges „kimerülése” és a résztvevőkben lejátszódó kognitív folyamatok azok, amelyek potenciálisan – és nem feltétlen intencionált módon – kihatással lehetnek az interakció „zenei atmoszférájára”, prosódiai szerkezetére is.

A disszertáció keretében megvalósított kísérletek között arra is fogunk példát találni, amikor nem lokális eseményekre vonatkozóan (mint amilyen a téma megváltozása), hanem az interakció egészét érintő globális sajátosságokra (pl. formális vagy informális keretek között zajlik-e egy beszélgetés) kérdezzük rá prosódiai jellemzők alapján. Hipotézisem szerint akkor bízhatunk az ehhez hasonló osztályzási feladatok sikerében, ha az akár bináris formában is feltehető kérdés (formális vagy informális) olyan egymástól élesen elkülönülő kategóriák zárt osztályára vonatkozik, amelyek reális eséllyel lehetnek hatással nem csak az interakcióban szereplő meg-

²³ Egy témaváltást detektáló algoritmus fejlesztése során tehát kétféle hibalehetőséggel is számolni kell. Témaváltást találunk ott, ahol nincs, vagy fordított esetben: elsiklunk felette.

nyilatkozások prozódijára, de például a beszélőváltás időbeli szerveződésére, ritmusára is. E tekintetben tehát a „prozódiát” a szokásosnál jóval tágabban, az interakció egésze felől, mint a résztvevők által konstruált, a beszélőváltás főtebb említett időszerkezetét is integráló globális egészet értelmezzük.

3. FEJEZET

A kutatás beszédtechnológiai háttere

A beszédtechnológia elsősorban műszaki tudományként számon tartott diszciplínája két nagyobb területre: a beszédjel analizálására és szintetizálására bontható fel. Mivel az utóbbi a disszertáció szempontjából nem releváns, az alábbi fejezetben az első terület a prozódia elemzése szempontjából sem megkerülhető feladataira fókuszálunk. Habár a feldolgozás folyamatának egyik fontos „végterméke” a beszédjelből előállított, további nyelvtechnológiai módszerek bemenetétül szolgáló szöveges átirat, az ehhez szükséges beszédfelismerő rendszerek tervezésével és működésével szintén nem foglalkozunk. A tényleges beszédfelismerést megelőző, egymással láncba szervezhető előfeldolgozási műveletek azonban esetünkben is hasonlóak, illetve bizonyos módszerek a beszéd szupraszegmentális szerkezetére is támaszkodhatnak:

1. A beszédaktivitás detektálása
2. A beszéd szegmentálása
3. Beszélő diarizáció
4. Az intonáció modellezése és lejegyzése

3.1. A beszéd detektálása és szegmentálása

A nemzetközi szakirodalomban *Voice Activity Detection* (Graf et al., 2015) vagy *Speech Activity Detection* (Sahidullah & Saha, 2012) terminusokkal jelölt eljárások alapvető feladata, hogy a bemenetként használt akusztikai jelben detektálják és lokalizálják a valóban emberi beszéd által gerjesztett, a további feldolgozási műveletek szempontjából érdekes szegmenseket. A detektálás történhet a jel adott időablakokra mért energiaszintjének folyamatos monitorozása vagy utólagos vizsgálata alapján (valós idejű és nem valós idejű alkalmazás), de a feladat korántsem

ennyire triviális olyan akusztikai környezetek esetén, ahol a beszélőkön kívül egyéb hangforrásokkal is számolni kell. Ezeket az egyéb hangforrások gerjesztette hangokat az eredő hullámforma spektrális tulajdonságai (vonalas vagy folytonos) szerint két nagyobb csoportba sorolhatjuk: zenei hangok és zajok. Az eljárás kimenete optimális esetben tehát egy, az akusztikai jel időtengelyének bizonyos szegmenseihez illesztett metaadat, aminek segítségével a beszédet izolálni tudjuk az összes többi lehetséges kategóriától (szünet, zene, zaj).

Hogy a szofisztikáltabb algoritmusokat ne terheljük indokolatlanul, első lépésként célszerű az energia-alapú megközelítésből kiindulni. A bejövő jelre vonatkozó első kérdés tehát így szól: „Történik-e bármi?”, amiről (a „bármiről”) majd egy következő fázisban lesz eldönthető, hogy ténylegesen beszédet takar-e. Feltételezve, hogy digitális jellel dolgozunk, az időablakban kapott minták (darabszámuk a mintavételi frekvencia, bitben meghatározható méretük a beállított kvantálás, numerikus formájuk a jel kódolásának függvénye) négyzetes átlagának (RMS érték) kiugró (egy bizonyos küszöbértéket meghaladó) változása adhat lehetséges támpontot ennek a „bárminek”, a jel aktivitásának észleléséhez, ahol a felhasznált küszöbértéket érdemes dinamikusán, a jel minimális energiaszintjéhez és SNR értékéhez (jel-zaj arány) igazítani, mivel még a tökéletes csend valószínűtlen állapotában is számítani kell arra, hogy a mérőmikrofon minimális alapzajjal dolgozik.

Az energiaalapú megközelítésre tekintsünk példaként három konkrét alkalmazást. Az elsővel a Kaldi beszédfelismerő rendszerben (Povey et al., 2011), a másodikkal a WebRTC projekt-re¹ épülő *voixen-vad*² node.js modulban, a harmadikkal pedig a Praat programban (Boersma & Weenink, 2016) találkozhatunk.

A Kaldi beszéd-detektáló moduljának³ elsődleges funkciója, hogy az egyes időablakokra mért MFCC (*Mel-frequency Cepstral Coefficients*) jellemzőket beszédaktivásra vonatkozó metaadatokkal lássa el a beszélő-felismerést, illetve a beszélőn és a csatornán belüli variancia kompenzációját szolgáló *I*-vektorok számításához, de egyszerűen felhasználható a hanganyag beszéd-szegmentumokra tagolásához is. Az eljárás kimenete az egyes időablakokra hozott bináris döntések sorozata, amelyek alapján a hanganyag (PCM kódolású, WAV formátumú fájl) szünetekkel tagolt, potenciálisan beszédet tartalmazó szegmentumokra bontatható fel. A Kaldi *github* repozitóriumában egy *Perl*-ben implementált, különféle paraméterekkel (a szünetek

¹<https://webrtc.org/>

²<https://github.com/voixen/voixen-vad>

³<https://github.com/kaldi-asr/kaldi/blob/master/src/ivectorbin/compute-vad.cc>

minimális, a beszédsegmentumok maximális hossza stb.) kalibrálható példát⁴ találhatunk a szegmentálás megvalósítására.

A *voixen-vad* node.js modul a Kaldi-val ellentétben valós időben érkező adatokkal is működik. Alapértelmezett esetben a webböngészőn keresztül, *float32* kódolásban (a minták ábrázolása 32 bites lebegőpontos számokkal történik), időben 20–30 milliszekundum hosszú adatcsomagokat dolgoz fel, az eredményeket pedig a node.js terminológiájában *event*nek⁵ nevezett, meghatározott függvények lefutását előidéző események (esetünkben: "VOICE" vagy "SILENT") formájában teszi lekérdezhetővé:

```
inputStream.on('data', function(chunk) {
    vad.processAudio(chunk, 160000, function(error, event) {
        if (event === VAD.EVENT_VOICE) {
            console.log("VOICE detected")
        }
        if (event === VAD.EVENT_SILENT) {
            console.log("SILENCE detected")
        }
    })
})
```

Az energia-szint ilyen kicsi ablakméret használata mellett természetesen állandó ingadozást fog mutatni, így folyamatos beszéd esetén is az alábbi esemény-minta ismétlődő szekvenciákra számíthatunk kimenetként: „VOICE”, „SILENT”, „SILENT”, „VOICE”. A nyers eredmények feldolgozásához (a beszédsegmentumok tényleges lokalizálásához) tehát be kell kalibrálnunk, hogy a „SILENT” címkével ellátott csomagok mekkora mennyisége indikálja a beszédsegmentum tényleges végét. Más szavakkal: be kell állítunk a beszédszünet minimális hosszát, amihez a csomagok mérete és a digitalizáláshoz használt mintavételi frekvencia szolgáltató technikai támpontot.⁶ Szintén célszerű beállítani egy minimális időintervallumot a potenciális beszédsegmentumokhoz, ami alatt a „VOICE” címkéjű csomagokat zajként ignoráljuk.

Az eddig tárgyalt energia-alapú módszerekhez hasonlóan a Praat program "Sound" típusú objektumokra alkalmazható, a dinamikus menüben „To TexGrid (silences)...” név alatt elérhető automatikus annotációja inkább tekinthető szünetdetektálásnak mint beszéd-detektálásnak. A

⁴<https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/utils/segmentation.pl>

⁵<https://nodejs.org/api/events.html>

⁶Ha például a mintavételezés 48000 Hertz a csomagméret pedig 4096 byte (1024 minta), akkor 10 csomag 213.33 milliszekundum időbeli hosszúnak felel meg (4096 x 8 / 32 / 48000 x 10).

kimenet a Praat esetében egy *TexGrid* formátumú átírat, ami a bemenetként kapott hangfelvételt „silent” és „sounding” címkékkel tagolja fel, így már az alapértelmezett címkeelnevezésből is látszik, hogy a „sounding” nem kizárólag beszéd lehet. Maga a módszer a hangfájl egészére számolt intenzitásadatok elemzésére támaszkodik, amit különféle paraméterekkel (decibelben megadott küszöbérték, minimális szünethossz stb.) szabályozhatunk. A kimenet utófeldolgozására itt csak abban az esetben van szükség (kész szegmentációt kapunk), ha „sounding” címkéjű szegmentumokat ténylegesen beszédre szeretnénk korlátozni.

Korábban említettük, hogy a zenei (periodikus) és a zörejszerű (nem-periodikus) hangok esetében a hullámforma *Fourier transzformációja* után kapott frekvenciaspektrum jellegzetes tulajdonságokat mutat, ami akár a szabály-alapú algoritmusok használatát is opcióként kínálja ahhoz, hogy az energia-alapú algoritmusok által detektált szegmenseket tovább kategorizáljuk. A beszéd spektrális tulajdonságai azonban jóval összetettebbek ennél. Vegyük alapul a beszédképzés gerjesztett szűrő modelljét (Fant, 1960). Míg bizonyos beszédhangoknál (magánhangzók, zöngés mássalhangzók) a hangképző üregek és a sugárzási impedancia átviteli függvénye által modulált kvázi-periodikus gerjesztés eredménye a zenei hangokra jellemző, diszkrét frekvencia-összetevőkre bontatható spektrum felé tendál, más hangoknál a lökés-hullámszerű (zöngétlen zárhangok) vagy a turbulens (zöngétlen réshangok) gerjesztés a spektrum folytonosságát fogja erősíteni. Természetesen ha nagyobb beszédegységeket, egész megnyilatkozásokat vizsgálunk (az energia-alapú detektálás eredményeként kapott szegmentumokat), a kváziperiodikus jelleg mindenképpen tetten érhető lesz, így közelítésnek elegendő, ha a kérdéses szegmensben leteszteljük, mérhető-e alapfrekvencia. Ezzel a köznapiban vett zajok nem-periodikus hullámformájú részét ki is zártuk, de a zenei hangokat és a fizikailag zenei hangnak számító zajokat (pl. autóduda, vonatfütyty) nem. Az eredmény pedig tovább rostálható, ha megvizsgáljuk, hogy a mért alapfrekvencia kívül esik-e a beszédre jellemző tartományon.

A szabály-alapú osztályozásnál megbízhatóbb eljárást képviselnek a gépi tanulást alkalmazó statisztikai megközelítések. A beszédfelismerésben is használt módszerekből kölcsönözve mind a három kategóriára (beszéd, zene, zaj) betaníthatunk egy akusztikai modellt, amivel a kérdéses szegmentumokat ütköztetni, majd kategorizálni lehet. Egy példaszekrényben⁷ ezt az utat kínálja fel a Kaldi keretrendszere is: az energia-alapú detektálás kimenetét egy a MUSAN korpuszon (Snyder et al., 2015) betanított GMM (*Gaussian Mixture Model*) modell segítségével bírálhatjuk felül. A eljárás hatékonysága tovább növelhető, ha a GMM modellek alapján ne-

⁷https://github.com/kaldi-asr/kaldi/tree/master/egs/bn_music_speech/v1

urális hálózatokat is betanítunk, a legtöbb azonban itt is a tanítóanyag méretén, a benne tárolt példák mennyiségén és változatosságán múlik.

A beszéd szegmentálásával kapcsolatban további kérdésként marad hátra, hogy mit határozzunk meg a szegmentálás alapegységének. Kiinduló helyzetben (a beszédaktivitás detektálása után) minden szegmentum beszédszünettől (ez technikailag lehet tényleges szünet, de zaj vagy zene is) a következő beszédszünetig tart. A felbontás a konkrét kutatási/alkalmazási cél függvényében tovább növelhető prozódiai frázisokra, illetve a beszélőváltások mentén történő beszédfordulókra bontással. Az előbbi természetesen nem előzheti meg frázisok előzetes modellezését és lejegyzését, az utóbbi pedig a következő szakasz témáját adó beszélő diarizáció kimenete lesz.

3.2. Beszélő diarizáció

A beszélő diarizáció⁸ (*Speaker Diarization*) a beszélő azonosításhoz (*Speaker Identification*) és a beszélő verifikációhoz (*Speaker Verification*) hasonlóan a beszélő felismerés (*Speaker Recognition*) összefoglaló területének egyik sajátos, az utóbbiakhoz képest kevésbé ismert, komplex alkalmazási típusa. A különbségek és hasonlóságok élesebb megvilágítása érdekében röviden mind három eljárást ismertetni fogjuk. Az általános nehézséget (ami az azonosításnál, a verifikációnál és a diarizálásnál is problémát jelent) az azonos beszélőhöz tartozó beszédjel nagy mértékű variabilitása okozza, amit két faktor: a használt csatorna és az egyes beszélőkön belül eleve meglévő variancia határoz meg. Ezzel a tényezővel még a szöveg-függő beszélőfelismerésnél (az egyedi sajátosságokat leíró modell alapjául szolgáló hanganyag a teszteléshez használttal azonos megnyilatkozást tartalmaz) is számolni kell, mivel a jel spektrális tulajdonságai ekkor sem mutatnak tökéletes egyezést. A DNS vagy az ujjlenyomat alapján történő azonosítás-sal/megkülönböztetéssel ellentétben beszéd alapján csak valószínűségi ítéleteket alkothatunk.

A beszélő azonosítás esetén rendelkezésünkre áll beszélőknek egy zárt halmaza (kriminalisztikai célú beszélőazonosításnál ez megegyezik a gyanúsítottak csoportjával) és egy ismeretlen beszélőtől származó hanganyag. A zárt halmazba tartozó beszélők hangjához készített modelleket egy, a konkrét alkalmazáshoz leginkább megfelelő beszédadatbázis alapján készített háttérmodellel (*Universal Background Model*) is kiegészítjük, az azonosítás során pedig arra a kérdésre keressük a választ, hogy mennyire valószínű, hogy az ismeretlen eredetű hanganyag valamelyik zárt halmazba tartozó beszélőtől származik és nem valaki mástól. A kriminalisztici-

⁸ A magyar nyelvű terminológiában Beke András munkája (Beke, 2015) nyomán *beszélő detektálásként* is honosult a kifejezés.

kai fonetikában ezt az úgynevezett hasonlósági arányszámmal (*Likelihood Ratio*) fejezik ki, ami pro és kontra két hipotézis *Bayes tétel* szerint meghatározható feltételes valószínűségére épül ($H1$: A hanganyag keletkezésének valószínűsége, ha az a kérdéses személytől származik, $H2$: A hanganyag keletkezésének valószínűsége, ha az nem a kérdéses személytől származik):

$$LR = H1 : H2 \quad (3.1)$$

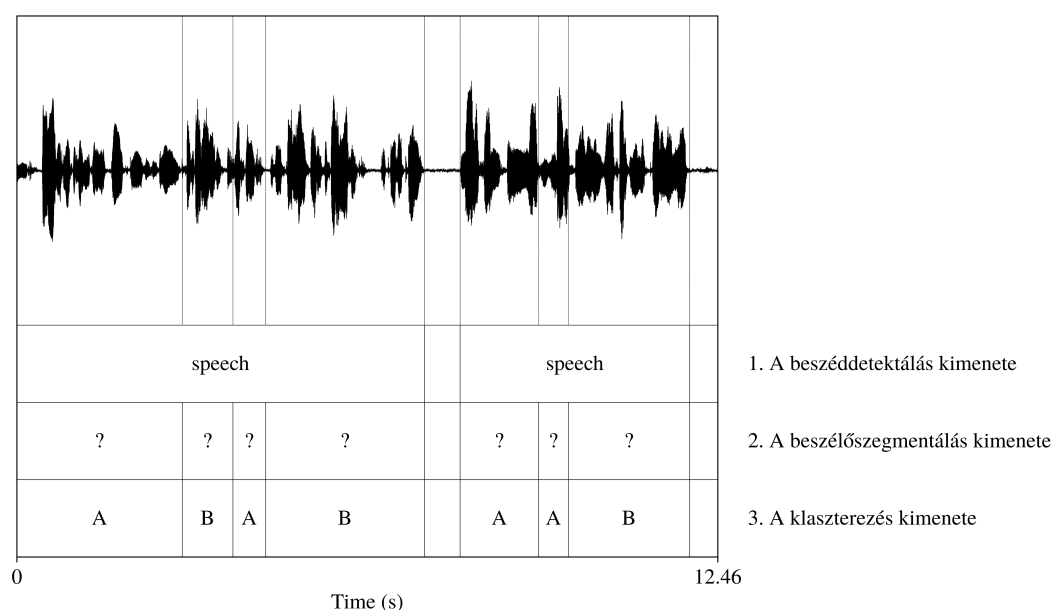
A gépi tanulással történő beszélőfelismerés egyik széles körben használt modern módszere (Ibrahim & Ramli, 2018; Kanagasundaram et al., 2011; Kanrar, 2017), hogy nagy mennyiségű, több ezer beszélőtől származó hanganyagból mért MFCC jellemzőket használva láncban betanítunk egy univerzális GMM háttérmodellt, majd egy másik számítási modellt (a Kaldi terminológiájában: *Ivector extractor model*), amivel később az egyedi MFCC jellemzőkből ki tudjuk nyerni a beszélőn belüli és a csatorna variabilitását is kifejező azonosítási vektorokat. Ezek az *I-vektorok* fogják leírni az azonosítandó beszélők egyedi modelljét, amit lineáris diszkriminancia analízis segítségével vethetünk össze az ismeretlen eredetű hangminta azonos módon számított vektoraival (a pontozáshoz használt LDA, illetve PLDA modellek elkészítése szintén betanítást igényel, ami az aktuális alkalmazásnak leginkább megfelelő hanganyagokból mért vektorokon történik).

A beszélő azonosítás és a beszélő verifikáció (vagy másik nevén: beszélő hitelesítés) alapvető módszereiben nem, csak a kimenetükben és kérdésfelvetésben különböznek. Míg az azonosításnál egy zárt halmazból választjuk ki a leginkább valószínűsíthető beszélőt, addig a verifikáció során a kérdéses hangmintát vagy elfogadjuk, vagy elutasítjuk a hitelesítendő beszélő modelljével való összevetés után. A verifikációt végző rendszer teljesítményének ellenőrzéséhez leggyakrabban használt mérőszám az EER (*Equal Error Rate*), aminek segítségével beállítható az a – pontozásnál használt – legoptimálisabb küszöbérték, ahol a téves azonosítások és a téves elutasítások aránya egyenlő.

A fentieket tovább komplikálva a beszélő diarizáció bemenete egy olyan hanganyag, amin egyszerre több, adott számú beszélő is hallható, a feladat pedig abban áll, hogy a felvételt a beszélőváltások mentén szegmentáljuk és minden egyes beszédsegmentumra vonatkozóan meghatározzuk, hogy az melyik beszélőhöz tartozik, mikor ki beszél a felvételen. Ez nem feltétlenül jelenti egyúttal a beszélők zárt halmazból kiválasztott beazonosítását. Diarizálásnak tekintjük azt is, ha beazonosítás nélkül izoláltuk a különböző beszélőkhöz tartozó szegmenseket, illetve meg tudjuk állapítani, hogy két vagy több szegmens (a felvétel különböző pontjain) azonos beszélőtől származik-e, függetlenül attól, hogy az illető kicsoda. Tehát az eljárás abban az alap-

értelmezett esetben is alkalmazható, ha felvételen megszólaló beszélőkhöz nem rendelkezünk előre elkészíthető modellekkel, sőt akár a beszélők száma is nyitott kérdés a vizsgálat elején.

Mint a 3.1. ábrán látható, a diarizáció első lépése a beszédsegmentumok előző szakaszban tárgyalt detektálása. A második lépést a beszélőváltások feltérképezése, a beszélőszegmentálás jelenti (*speaker change detection*). Szerencsés esetben (lásd a 4. és 5. címkét a 3.1. ábra 3. annotációs szintjén) a beszélőváltások beszédsszünetek mentén történnek (a szünetalapú beszélőszegmentálók így az elsővel a második lépést is teljesítik), azonban spontán informális beszélgetéseknél gyakran előfordul, hogy egy adott beszédsegmentumon belül (két szünet között) több beszélő váltja egymást. A feladatot tovább komplikálják az úgynevezett átfedő beszéddek, amikor két vagy több beszélő megszólalásának van közös időbeli metszete. A beszélőváltás szempontjából releváns változások detektálására kifejlesztett algoritmusok közül a metrikus alapú, a Bayes-féle információs kritériumot és különféle akusztikai jellemzőket (MFCC, alaphangmagasság, az energiaspektrum temporális stabilitása, a spektrális burkológörbe alakja stb.) használó szegmentálók a legnépszerűbbek (Beke, 2015).



3.1. ábra. A beszélő diarizáció lépései a beszéd-detektálástól a beszélők klaszterezéséig.

A szegmentálást követő utolsó lépése az úgynevezett klaszterezés, a beszélőszegmentálás kimenetében lévő szegmentumok homogén csoportokhoz (klaszterekhez) való rendelése (a 3.1. ábra 3. annotációs szintjén ezeket a különböző beszélőket reprezentáló csoportokat A és B címkével jelöltük), amelynek végén az egy csoportba sorolt szegmentumokat azonos címkével látjuk el, hogy a kimenetben pontos képet kapjunk a beszélők számáról, illetve megnyilatkozásaik pozíciójáról. A klaszterező eljárásokban leginkább az alulról felfelé egyesítő hierarchikus

technikákat preferálják (Anguera et al., 2006; Beke, 2015), ahol a kiinduló hipotézis szerint a beszélőszegmentálás eredményeként nyert összes szegmentumot külön beszélő-klaszternek tekintjük, majd iteratívan haladva addig egyesítjük őket, amíg egy úgynevezett megállási kritériumnak megfelelően el nem érjük a beszélők valós számát. A klaszterezéshez jellemzőként felhasználhatjuk a beszélő azonosítás kapcsán korábban már említett I-vektorokat (Zajíc et al., 2016) is. A Kaldi *github* repozitóriumában erre szolgál példaként: a *callhome diarization* nevű recept⁹ első verziója, de újabban az X-vektorokon alapuló, neurális hálózatokkal kiegészített megoldások (Snyder et al., 2018; Snyder et al., 2019) is egyre népszerűbbek.

A eljárást megvalósító automatizált rendszerek teljesítményének mérésére leginkább a DER (*Diarization Error Rate*) mérőszámhoz definiált metrikát szokták alkalmazni, de léteznek alternatív megközelítések is, amelyek például az átfedő beszédek jelenlétével is kalkulálnak (Galibert, 2013). A DER sztenderd metrikája az alábbi, a beszédaktivitás detektálásának sikerességét is ellenőrző hibalehetőségekkel kalkulál: a tévesen beszédként azonosított szegmensek hossza (FA), a detektálás során elmulasztott beszédsegmentumok hossza (MISS), a hibásan klaszterezett, nem megfelelő beszélőcímkével ellátott szegmensek hossza (CONF). A DER hibaarány értékét pedig úgy kapjuk meg, ha ezek összegét elosztjuk a referenciaanyagban beszédként címkézett szegmensek hosszának összegével (TOTAL):

$$DER = \frac{FA + MISS + CONF}{TOTAL} \quad (3.2)$$

A beszélő diarizáció automatizált kivitelezésére a nemzetközi szakirodalomban és a kapcsolódó internetes repozitóriumban számos konkrét, beszédadatbázisokon tesztelhető, nyílt hozzáférésű implementációval találkozhatunk (Vozáriková & Juhár, 2015). A Kaldi-hoz hasonlóan I-vektoron alapuló megoldásokat támogat az ALIZE (Bonastre et al., 2005), az egyéb megközelítések közül e helyütt csak két példát a *Javában* implementált LIUM (Rouvier et al., 2013) és az e-magyar projektben (Váradai et al., 2017; Kornai & Szekrényes, 2017) is felhasznált SHOUT (Huijbregts, 2008) nevű szoftvert emeljük ki. Hazai viszonylatban e területen a már többször hivatkozott Beke András végzett figyelemreméltó és úttörő munkát (Beke, 2015).

3.3. Az intonáció modellezése és lejegyzése

Az intonáció modellezését eltérő szempontokból megközelítő fonológiai paradigmákat kiszolgáló több olyan eljárás is napvilágot látott, amelyek az élőbeszéd prozódiai szerkezetének

⁹https://github.com/kaldi-asr/kaldi/blob/master/egs/callhome_diarization/

lejegyzését hivatottak sztenderdizálni, illetve a lehetőséghez mérten automatizálni. Az alábbi fejezetekben három, az intonáció szimbolikus lejegyzésére (ToBI), parametrizált modellezésére (Tilt), illetve pszichofizikai alapú stilizálására (Prosogram) szolgáló eljárás kerül bemutatásra. Habár a prozódiai információk feldolgozására a beszédtechnológia területéről számos egyéb példát találhatunk, a kidolgozott módszerek közül itt most csak azokat ismertetem, amelyek három – a disszertáció szempontjából fontos – kritériumoknak megfelelően: (1) olyan kimenettel rendelkeznek, amely explicit formában reprezentálja, nyelvészeti szempontból elemezhetővé teszi az élőbeszéd prozódiai szerkezetét, (2) publikusan elérhető és szabadon felhasználhatók, (3) a disszertációban keretében megvalósított fejlesztésekre jelentősebb hatással voltak. Ugyanakkor fontosnak érzem felhívni a figyelmet Reichel et al. (2014) magyar vonatkozású kutatására is, ami saját fejlesztéseimhez hasonlóan, az alapprofrendencia stilizálása útján – az intonáció megnyilatkozás-szintű karakterdallamainak (Varga, 1994) automatikus címkézésével kapcsolatban ért el fontos eredményeket.

3.3.1. ToBI

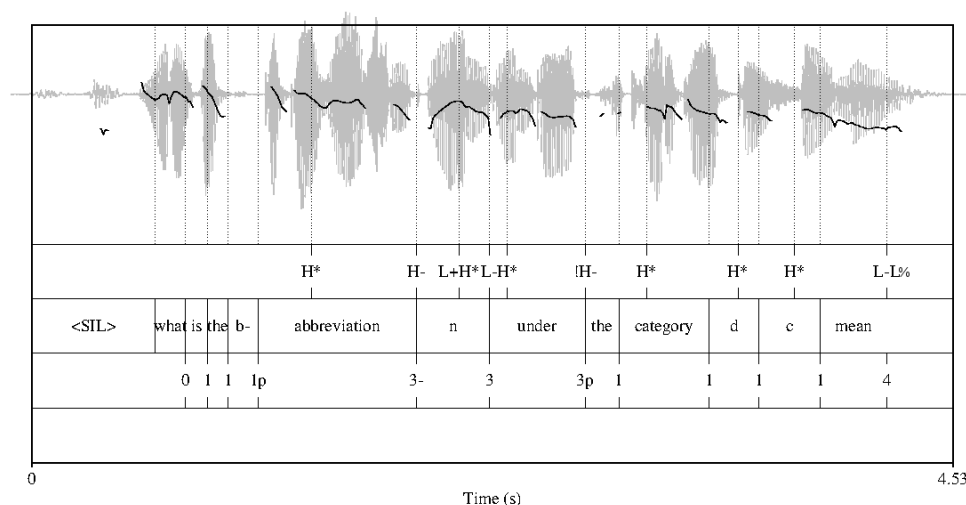
A ToBI („Tones and Break Indices”) egy eredetileg az amerikai angol nyelvere kidolgozott (Silverman et al., 1992), de később jóval szélesebb körben elterjedő¹⁰, a prozódia szimbolikus lejegyzésére szolgáló annotációs, az egyes konkrét nyelvek sajátosságaira specifikálható keretrendszer. Habár a címkézés automatizálásra is történtek már kísérletek (Rosenberg, 2010), a projekt alapvetően nem egy számítógépes alkalmazást, hanem egy – az autoszegmentális fonológia elvein (Pierrehumbert & Hirschberg, 1990) alapuló – intonációs modell szimbólumrendszerét, mint könnyen elsajátítható lejegyzési protokollt igyekszik népszerűsíteni. Az elemzés tényleges kivitelezése történhet valamilyen általunk preferált, az alapprofrendencia mérésére és vizualizálására, illetve címkézésre is alkalmas szoftveres környezetben (ez lehet például a Praat vagy a KTH által fejlesztett WaveSurfer¹¹ program).

A ToBI rendszerében a beszéd elemzése az autoszegmentális fonológia elveinek megfelelően több párhuzamos szinten történik, amelyek közül az úgynevezett „tone tier” és a „break-index tier” fogja reprezentálni a megnyilatkozások tényleges prozódiai szerkezetét. Az „orthographic tier” a beszéd ortografikus lejegyzésére, a „miscellaneous tier” pedig egyéb, közvetlenül nem az

¹⁰ A 1990-es évek elején induló projekt hivatalos honlapja alapján jelenleg az alábbi nyelvekre történt komplett implementáció: angol (amerikai, brit, glasgow-i dialektus), német, japán, koreai. Előrehaladott fejlesztések folynak az alábbi nyelvekre: görög, katalán, portugál, szerb, mandarin, kantoni, spanyol, tajvani.

¹¹ <http://www.speech.kth.se/wavesurfer/>

intonációra vonatkozó, de annak szerkezetét befolyásolható jelenségek (pl. köhögés) jelölésére használatos.



3.2. ábra. Címkézés a ToBI rendszerében. Az elemzésre szolgáló annotációs szintek felülről lefelé haladva: „tone tier”, „ortographic tier”, „break-index tier”. Az ábra a Praathoz letölthető ToBI példák (<http://www.cs.columbia.edu/>) felhasználásával készült.

A „tone tier” az alapprofrekvencia görbén megfigyelhető, de a fonológiai elemzés irányából is relevanciával bíró (perceptuálisan disztinktív) intonációs események szimbolikus lejegyzését tartalmazza, amelyek a Tobi rendszerében két fő csoportra, az úgynevezett dallamhangsúlyokra („pitch accent”) és határjelző tónusokra („boundary tones”) oszthatók fel. A dallamhangsúlyok jelölésére kétfajta, az aktuális fonetikai megvalósulást jellemző céltónus (a lokális F0 maximumot jelentő „(H)igh” és a lokális F0 minimumot jelentő „(L)ow”) szolgál, amelyek vezető (pl. „L+”) vagy követő (pl. „+H”) tónusokkal is kombinálhatók. Mint a 3.2. ábrán látható, a dallamhangsúlyok releváns tartományát mindig az elődleges szóhangsúly szótagja jelenti, a megnyilatkozás nem hangsúlyos elemeinek F0-kontúrjai csak akkor képezik elemzés tárgyát, ha valamiért relevánsak.

A határjelző tónusok lejegyzése a „break-index” szint szerkezetével szoros összefüggésben áll. Az utóbbi szint szegmentálja ugyanis a beszédet úgynevezett prozódiai frázisokra a szavak közötti absztrakt szóhatárként értelmezendő „szünetek” öt szintű (0–4) indexelésén keresztül, ahol az 1. index a szóhatárok frázishatárt nem képző alapesete. Az egymással összemosódó szavak határait (pl. „Who is -> Who’s”, „What is -> What’s”) a 0. index jelzi. A 2. index azokra az esetekre érvényes, ahol a szóhatár az intonációra nincs hatással, vagy ha mégis, nincs érzékelhető szünet. A 3. index választja szét az alacsonyabb szintű, köztes frázisokat („intermediate phrase”), amelyek határa a „tone” szinten egy frázis tónussal („L-” vagy „H-”) jelölhető. Vé-

gúl a 4. index, ami egy teljes intonációs frázis határát jelölve a „tone” szinten valamilyen záró tónussal („final boundary tone”) terminál.

A teljes intonációs frázisoknak három tipikus, az angol nyelvű megnyilatkozások értelmezése szempontjából fontos esetét különböztethetjük meg:

1. Az L-L% kontúrral megvalósuló deklaratív frázis, amely egy befejezett kijelentés modalitásának feleltethető meg
2. Az L-H% kontúr az előzővel ellentétben a közlés befejezetlenségre, a folytatás szándékára utal (általában ezzel a kontúrral valósulnak meg a felsorolások elemei)
3. A H-H% kontúr az eldöntendő kérdések intonációs formája

Mivel a ToBI lényegét tekintve egy keretrendszer, az angoltól eltérő nyelvek prozódijának a leírására a rendszer azon elemeit használják, amelyek az adott nyelvben relevánsak. Így az egy bizonyos nyelvre kidolgozott implementáció nem feltétlenül lesz változatlan formában alkalmas egy másik nyelv automatikus elemzésére. Saját kutatásaimhoz egy olyan alkalmazás fejlesztését tűztem ki célul, ami a nyelvek szuprasegmentális fonológiai sajátosságait figyelmen kívül hagyva csak az empirikusan megragadható prozódiai eseményekre fókuszál. Ugyanakkor saját rendszerem annotációs címkéi a ToBI lokális alaphangmagasságokat jelölő kategóriáihoz hasonlóan igyekeznek ezeket az eseményeket az adott beszélő relatív hangterjedelmében lokalizálni.

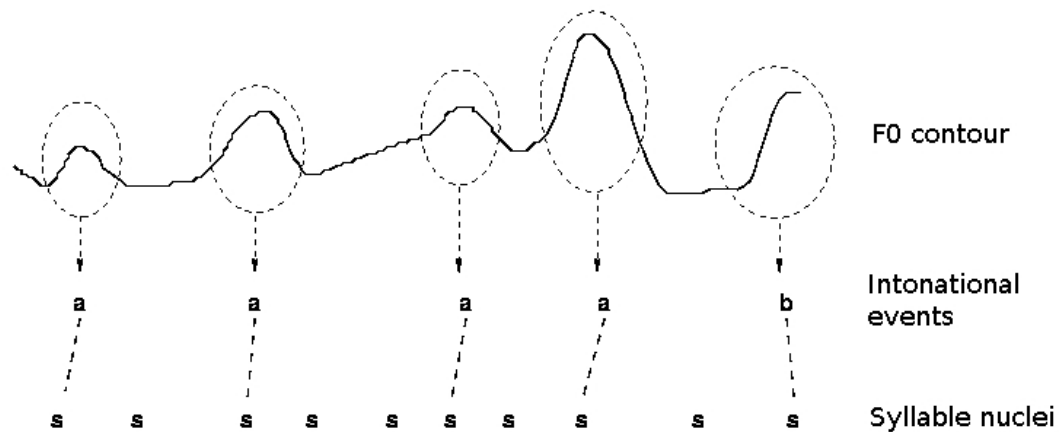
3.3.2. Tilt

Az alaphangmagasság-kontúr fonetikai szempontból való elemzésére, illetve szintetizálására is felhasználható Tilt modell¹² a beszéd megnyilatkozásait úgynevezett intonációs események sorozatára tagolja fel, amelyeket nem kategorikus címkékkel, hanem folytonos paraméterekkel jellemez (Taylor, 2000). Az események detektálása egy rejtett Markov modelleken (HMM) alapuló, a HTK keretrendszerben¹³ implementált algoritmus segítségével történik, amit előzetesen egy manuálisan szegmentált korpuszon tanítottak be. Mint a 3.3. ábrán látható, minden intonációs esemény az alaphangmagasság-kontúr fonetikailag megragadható változásaihoz, a beszéd-szegmentumok szintjén pedig egy-egy szótaghoz kötődik. A ToBI rendszerével rokonítható tulajdonság, hogy az események a Tilt modellben is két csoportra, az ábrán *a* karakterrel jelölt

¹²http://festvox.org/docs/speech_tools-1.2.0/c16909.htm

¹³<http://htk.eng.cam.ac.uk/>

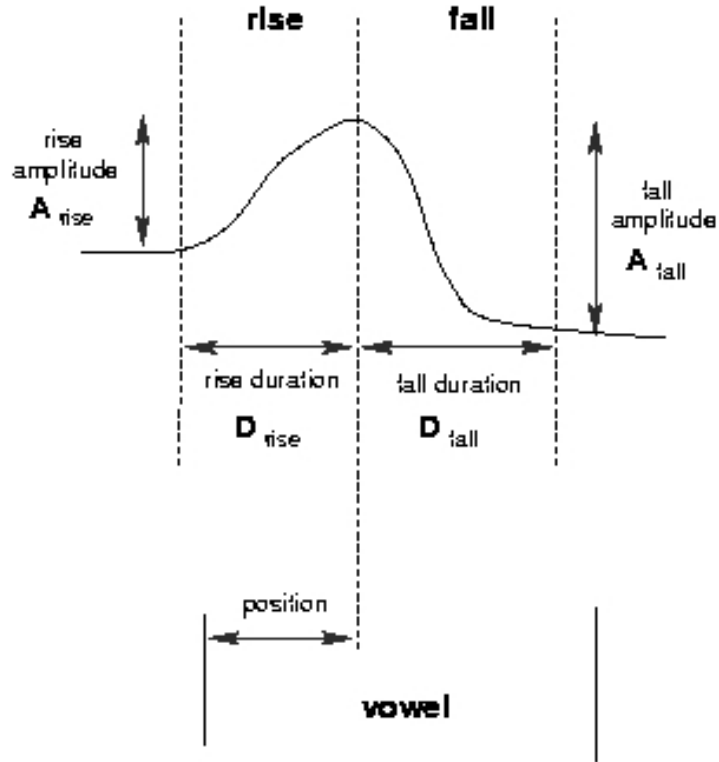
dallamhangsúlyokra („pitch accents”) és a *b* karakterrel jelölt határjelző tónusokra („boundary tones”) oszthatók fel. Az utóbbiak a Tilt-ben mindig emelkedő tónussal valósulnak meg.



3.3. ábra. Az Tilt modell intonációs eseményeinek ábrázolása az F0-kontúrral és a szótagtengellyel való összefüggésükben. Forrás: http://festvox.org/docs/speech_tools-1.2.0/c16909.htm

Az intonációs események a végleges, Tilt-specifikus reprezentációt megelőzően az RFC („rise/fall/connection”) modell paramétereit alapján kerülnek elemzésre, ami minden eseményt egy emelkedő („rise”) és egy eső („fall”) kontúrból felépülő karakteregyüttesként ragad meg. Mint a 3.4. ábrán látható, az események RFC-alapú leírása 6 paraméter segítségével végezhető el:

1. Az emelkedő szakasz Hertzben kifejezett amplitúdója (A_{rise}).
2. Az eső szakasz Hertzben kifejezett amplitúdója (A_{fall}).
3. Az emelkedő szakasz időtartama (D_{rise}).
4. Az eső szakasz időtartama (D_{fall}).
5. Az esemény Hertzben kifejezett csúcspontja.
6. A csúcspont szótagon belül elfoglalt pozíciója.



3.4. ábra. Az RFC modell paramétereit. Forrás: http://festvox.org/docs/speech_tools-1.2.0/c16909.htm

Az RFC séma tetszőleges fonetikai forma lejegyzésére alkalmazható: ha az eseményből valamelyik összetevő hiányzik, akkor az emelkedésre vagy esésre vonatkozó paraméterek nulla értéket vesznek fel. Az eseményekhez társított végleges reprezentáció a hat paraméterből explicit formában csak az utolsó kettőt őrzi meg, az emelkedő és az eső szakasz attribútumait egyetlen, azok (a képletben A -val jelölt) amplitúdója és időtartama (D) alapján kiszámított *tilt* jellemzővel fejezi ki:

$$tilt = \frac{|A_{rise}| - |A_{fall}|}{2(|A_{rise}| + |A_{fall}|)} + \frac{D_{rise} - D_{fall}}{2(D_{rise} + D_{fall})} \quad (3.3)$$

A tilt ezen összetett jellemzője mellett két további egyszerűbb mutatót különböztetnek meg az esemény-komponensek (emelkedés és esés) amplitúdójának (A_{event}) és időtartamának (D_{event}) összegeként:

$$A_{event} = |A_{rise}| + |A_{fall}| \quad (3.4)$$

$$D_{event} = D_{rise} + D_{fall} \quad (3.5)$$

A *tilt* paraméter és ez utóbbi kettő (A_{event} és D_{event}) ismeretében az RCA paraméterek (lásd 3.6. és 3.7), illetve az eredeti F0-görbe stilizált verziója is rekonstruálható.

$$A_{rise} = \frac{A_{event} + (1 + tilt)}{2} \quad (3.6)$$

$$A_{fall} = \frac{A_{event} + (1 - tilt)}{2} \quad (3.7)$$

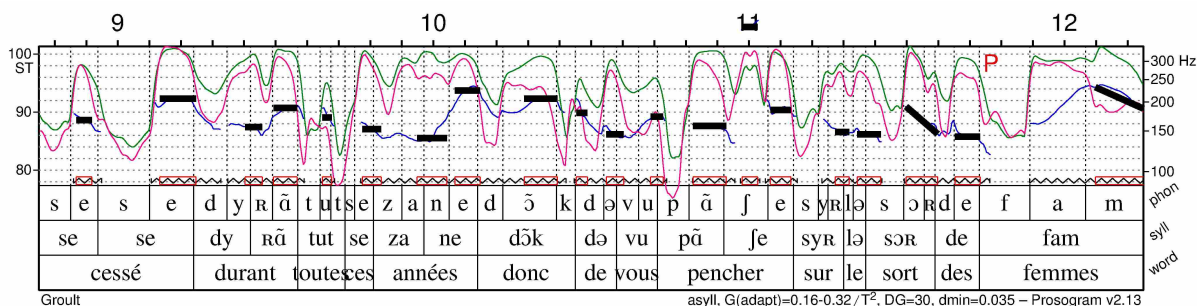
A saját kutatásaimban a Tilt parametrikus megközelítése helyett olyan kategorikus esemény-típusokat alkalmaztam, ahol az emelkedés és az esés szakaszai egymást követő, különálló eseményekként őrződnek meg, amelyek detektálása a Tilt-tel szemben nem igényli a beszédjel előzetes, tanítóanyagot igénylő szegmentálását. Viszont a Tilt-hez hasonlóan az intonációs események intervallumait az RFC modell paraméterei alapján elemzett alapfrekvencia kontúrhoz társítom hozzá, a beszéd szegmentális egységeitől és egyéb jellemzőitől (pl. intenzitás, hang-súly) függetlenül.

3.3.3. Prosogram

A Prosogram egy Piet Mertens (Mertens, 2004) által fejlesztett, a Praat program szkriptnyelvén implementált, szabályalapú alkalmazás, ami az alapfrekvencia-kontúr perceptuálisan releváns változásainak stilizálására szolgál. Az eljárás egy korábbi, a tonális érzékelés pszicho-akusztikai modelljét (Hart, 1976) követő módszeren (d'Alessandro & Mertens, 1994) alapul, ami a géppel mért alapfrekvencia-értékeket szótagszinten integrálva a Glissando küszöb (G_{tr}) segítségével határozza meg, hogy a kérdéses dallamszekvencia a hangmagasság érzékelhető változása alapján milyen formájú (emelkedő, eső vagy szinttartó) tónusokra tagolható tovább. A 3.8. képletben a T értéke a szegmentum – a G_{tr} küszöbvel fordítottan arányos – időtartamát fejezi ki: minél hosszabb a szegmentum, annál könnyebben érzékelhető a hangmagasság változása.

$$G_{tr} = \frac{0.16}{T^2} \quad (3.8)$$

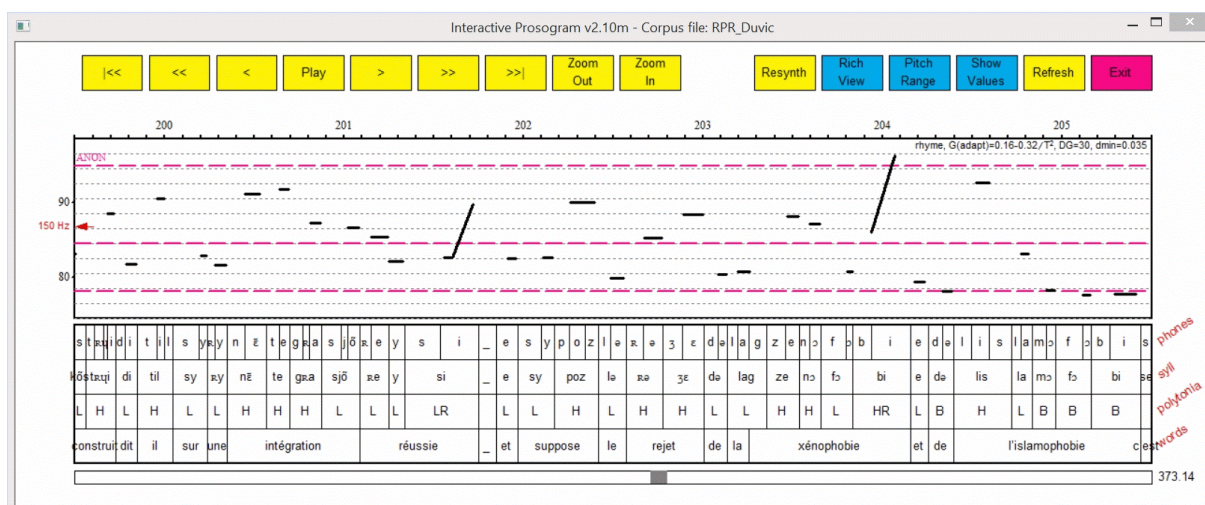
Mint a Prosogram grafikus kimenetét szemléltető 3.5. ábrán látható, a beszéddallam vastag fekete vonalakkal jelölt stilizációi, amelyek a második szegmentális tengelyen (*syll*) mindig egy-egy szótaggal társíthatók össze, az intonáció intraszillabikus változásait reprezentálják. A stilizálás támaszkodhat a szótaghatárokat tartalmazó manuális annotáció szegmentumaira, vagy ha ilyen a szótag-szintű átirat nem áll rendelkezésre, akkor az intenzitás-görbe alapján a szótagokra szegmentálás automatikusan is kivitelezhető.



3.5. ábra. A Prosogram grafikus kimenete. Kék görbe: F0-kontúr. Vastag fekete vonalak: stilizált intonáció. Zöld és rózsaszín görbe: intenzitáskontúr. Az alsó három szint a megnyilatkozás manuálisan annotált szegmentális elemeit (beszédhangok, szótagok, szavak) jeleníti meg. Forrás: <https://sites.google.com/site/prosogram/>

A PDF, EPS és PNG formátumban elérhető grafikus kimeneten kívül a Prosogram lehetőséget ad a stilizációk Praat PitchTier formátumban való elmentésére is, ami a dallamszegmentumok kezdő és végpontját idő-frekvencia értékpárok formájában tárolja. Ezek a „köztes adatok” további automatikus feldolgozás tárgyát képezhetik, illetve a Praat és az ELAN programokban interaktív üzemmódban is elemezhetők.

A Prosogram kimenetének kategorikus jellemzőkkel való felcímkezésére *Polytonia* néven Mertens egy automatikusan implementálható transzkripció szabványt is kidolgozott, ami a stilizációkat a beszélő 5 szintre felosztott hangterjedelmében (*Bottom, Low, Middle, High, Top*) elfoglalt pozíciójuk és formájuk (*Rise, Fall*) alapján jegyzi le. A 3.6. ábrán látható szaggatott vonalak ezeknek a hangmagasságszinteknek a határait szemléltetik. A harmadik annotációs szinten (*polytonia*) látható címkék ezeket a stilizált szegmentumokhoz társított szintértékeket a kontúr formájával kiegészítve (de a „szinttartó” kategória alul-specifikált eseteit jelöletlenül hagyva) írják le a megnyilatkozás intonációjának fonetikai szerkezetét.



3.6. ábra. A Prosogram kimenete a Praat program interaktív felületén. Forrás: <https://sites.google.com/site/prosogram/>

Saját kutatásaim sok tekintetben támaszkodtak a stilizálás Prosogramban is alkalmazott elvére. Lényeges különbség azonban, hogy a Glissando küszöb szótag-szintű alkalmazásával nyert dallamszegmentumok jellemzése helyett, az intonáció holisztikusabb, akár több szótagon is átívelő elemzésére törekedtem, ami feltevésem szerint jobban leképezi a percepcióban megjelenő hanglejtés-egységeket.

4. FEJEZET

A kutatás anyaga és módszertani eredményei

Az alábbi fejezet az értekezés keretében folytatott, a prozódia automatikus annotációját célzó fejlesztői tevékenységem módszertanát taglalja. Elsőként az annotáció kimenetét is tartalmazó *HuComTech* korpuszt mutatom be, amelyet nem csak a fejlesztéshez használt algoritmusok tesztelésére, de a hipotézisek igazolására kidolgozott, a 6. fejezetben szereplő kísérletekben is felhasználtam. A fejezet címében az „eredmény” arra utal, hogy az itt kidolgozott és implementált módszerek nem csak eszközök voltak a kutatás szolgálatában, de egyben annak közvetlen céljai is.

4.1. A kutatás anyaga

A kutatás elsődleges empirikus adatforrását és tesztanyagát jelentő, 50 órányi spontán beszédet tartalmazó *HuComTech* multimodális élőnyelvi korpusz (Hunyadi et al., 2012) létrehozásában és fejlesztésében egészen a kezdetektől aktívan közreműködtem.¹ A korpusz első közzétételére 2014-ben került sor a Language Archive webes felületén², de az új elemzési szempontok és módszerek integrálásával az anyag azóta is folyamatosan bővül. A korpusz magját azok a 10–20 perc hosszúságú interjúfelvételek alkotják, amelyek 111 adatközlő (54 nő és 67 férfi, átlagéletkor: 22 év, többségében egyetemi hallgatók) részvételével 2010 tavaszán készültek a Debreceni Egyetemen a TÁMOP 4.2.2-08/1-2008-0009 azonosító számú, „Az ember–gép kommunikáció

¹ Legfontosabb feladataim a korpusz (hamarosan bemutatásra kerülő) felvételeinek elkészítése (interjúztatás és technikai felügyelet), az annotáció folyamatának koordinálása és technikai támogatása (pl. automatikus ellenőrzések), a prozódiai annotáció kialakítása, végül az annotációk és a felvételek feldolgozása (szinkronizálás, összefűzés, konvertálás, metaadatok elkészítése stb.) volt.

² <https://hdl.handle.net/1839/00-0000-0000-001A-E17C-1@view>

technológiájának elméleti alapjai” című pályázat támogatásával. A szövegfelolvasásos feladatokat követő interjúk minden alany esetén két eltérő scenáriót, egy szimulált állásinterjú és egy azt követő informális beszélgetés kereteit követték. A társalgások utólag összeszinkronizált hang- és képanyaga – a felvételek minőségére és a különböző modalitások izolált feldolgozhatóságára való tekintettel – párhuzamosan, külön csatornákon rögzítettünk. Az elkészült nyersanyaghoz később társultak hozzá a korpusz legértékesebb és legidőigényesebb részét jelentő annotációk (Pápay et al., 2011), amelyek a felvételek különböző szempontokból történő elemzésének eredményét szöveges, további számítógépes feldolgozásra alkalmas címkék formájában tárolják.³ Bár a hang és videó felvételek elemzése eltérő szoftveres környezetben történt (hang: Praat, videó: Qannot⁴), az annotációk összefűzése és konvertálása során egy XML alapú, a multimodális korpuszok esetén sztenderdnek számító, az ELAN annotáló és lekérdező szoftver (Wittenburg et al., 2006) által használt EAF formátumban egységsítettünk minden anyagot.

A korpusz létrehozásának elsődleges célja természetesen nem a disszertáció céljaként korábban megfogalmazott hipotézisek ellenőrzése⁵, hanem a kommunikáció formális, egy általános alapszerkezetből kiinduló, majd ezen alapszerkezet funkcionális és pragmatikai kiterjesztésével a kommunikáció egyedi reprezentációit, eseményeit levezető generatív modelljének⁶ (Hunyadi, 2011) verifikációjához és technológiai alkalmazásához szükséges empirikus erőforrás létrehozása volt. A korpusz annotálása során felhasznált elemzési sémák viszont kellően alulspecifikáltak és a lehetőségekhez mérten elméletsemlegesek⁷ ahhoz, hogy a korpusz felhasználását egyéb (akár a modell hatókörén kívüli) kutatási célok számára is lehetővé tegyék.⁸ A disszertáció tekintetében leginkább érdekes részletek kiemelése előtt fontosnak tartom, hogy a létrehozott

³ Mivel nem szövegek, hanem időbeli események annotációjáról van szó, a „tárolás” egyben az annotációs címke időbeli rögzítését is jelenti az elemzés alapjául szolgáló médiaanyag időtengelyén.

⁴ A Qannot videó annotáló program fejlesztése szintén a TÁMOP 4.2.2-08/1-2008-0009 projekt keretében történt a Debreceni Egyetemen. Jelenleg csak belső használatra szánt dokumentációja elérhető.

⁵ A korpusz megtervezése és a magját alkotó interjúfelvételek elkészítése már az előző szakaszban bemutatott hipotézisek és kutatási célok megszületése előtt, 2010 nyaráig lezajlott.

⁶ Ahogyan a generatív grammatikában lehetséges mondatokról, ebben modellben a kommunikáció lehetséges eseményeiről, eseménysorozatairól beszélünk, amelyek a modellben bemutatott levezetési szabályok és alapelemek („primitívek”) segítségével generálhatók.

⁷ A pragmatikai annotáció esetében például nem volt megkerülhető, hogy a címkézés során használt kategóriák ne támaszkodjanak a beszédaktus-elmélet egy – a multimodalitás igénye alapján – kiterjesztett változatára.

⁸ Az adatok „elmélet-semlegessége” és „alul-specifikáltsága” más szempontból a modell verifikálását is szolgálják, hiszen ha a modell alapján kerültek volna kidolgozásra, nem szolgálhatnának az elmélet bizonyításának plauzibilis alapjául.

adathalmaz tulajdonságairól átfogó, az annotációs rendszert egészében bemutató, módszertani szempontból is részletes képet adjak.

4.1.1. A felvételek

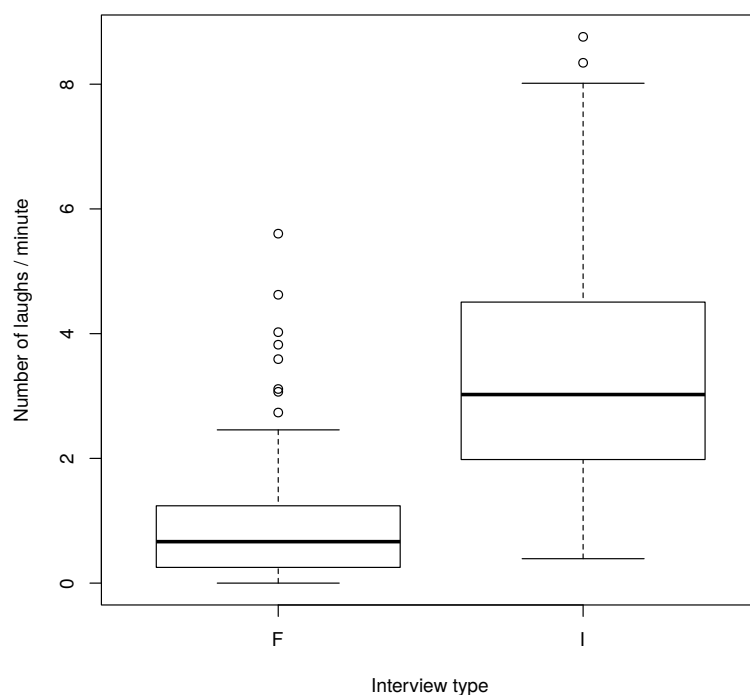
A felvételek az Debreceni Egyetem Angol-Amerikai Intézetének stúdiójában készültek. A résztvevők a kezelőpulttól hangszigetelt gipszkarton falakkal leválasztott felvevőhelyiségben egymással szemben, ülő pozícióban foglaltak helyet. A helyiségben kettő darab Sony HDRXR520VE típusú kamera rögzítette az adatközlőt, egy pedig az interjúvezetőt (lásd 4.1. ábra). A interakció hanganyaga kettő darab, a résztvevőkre irányított Shure 16A típusú mikrofon segítségével 44100 Hz mintavételezési frekvencián és 16 bites kvantálással, 2 csatornán került rögzítésre.



4.1. ábra. Az adatközlőt (fent) és az interjúvezetőt (lent) rögzítő kamerák képanyagát azonos időpontban bemutató pillanatkép. Felvevőeszközök elhelyezkedése: (1.) az interjúvezetőt rögzítő kamera (2.) az adatközlő hangját rögzítő mikrofon (3-4.) az adatközlőt rögzítő kamerák (5.) az interjúvezető hangját rögzítő mikrofon.

A résztvevők első feladata 20 mondat és 6 szó felolvasásából állt, amelyek egy, a BME Fonetikai Laborja által rendelkezésünkre bocsátott, változatos hangkapcsolatokat reprezentáló szövegtörzsből származtak. A felolvasásukról készült felvételek a későbbiekben beszédfelismerő rendszerek tanítóanyagaként hasznosultak.

A felolvasásokat követő szimulált állásinterjú afféle „próbainterjúként” volt tálalva az adatközlők számára, amely során lehetőségük volt kipróbálni magukat néhány, az éles állásinterjúkon tipikusan elhangzó, a jelölt tanulmányaira, céljaira, igényeire, motiváltságára és önismertetére vonatkozó általános kérdés kerestüzeiben. A formális keretek megőrzése érdekében és a beleélést segítő az interjúalanyok kopogást követően léptek be a felvételihelyiségbe, ahol a hivatalos bemutatkozás után magázó hangnemben válaszoltak az előre kidolgozott interjúkérdésekre. Az informális dialógus kiindulópontját az állásinterjú szerzett benyomások megbeszélése jelentette („Na, milyennek érezted?”), ami után a társalgás menete egy személyes témákat („Mesélj egy pozitív élményről, ami mostanában ért!”, „Mesélj a barátaidról” stb.) boncolgató kérdéssor mentén haladt tovább. A motiválatlan témaváltások miatt az interjúvezető ebben az új, közvetlenebb hangvételi szituációban is megőrizte irányító szerepét, az adatközlők viszont – a szimulált állásinterjú feszültebb helyzete alól felszabadulva – sokkal oldottabb viselkedést tanúsítottak. Bizonyos esetekben egészen meglepő nyíltsággal tárták fel személyes életük intimebb részleteit (családi tragédiák, párkapcsolati problémák stb.) is. Ezt a különbséget szemlélteti az interjúban előforduló nevetések percenkénti előfordulását ábrázoló 4.2. ábra. A formális interjú okozta stresszhelyzet megszűnése mellett ez a váratlan feloldódás az interjúvezetővel kialakult speciális viszonytal is magyarázható. Az állásinterjú után már nem két egymásnak teljesen idegen személy lépett kényszeredett interakcióba egymással minden természetes apropó nélkül. A közösen átélte szimuláció ugyanis referenciapontként szolgált a konverzáció gördülékeny folytatásához, amiben a kérdező fél hatalmi pozíciójából látszólag kilépve, egyenrangú partnerként adott lehetőséget a kötetlen feltárulkozásra.

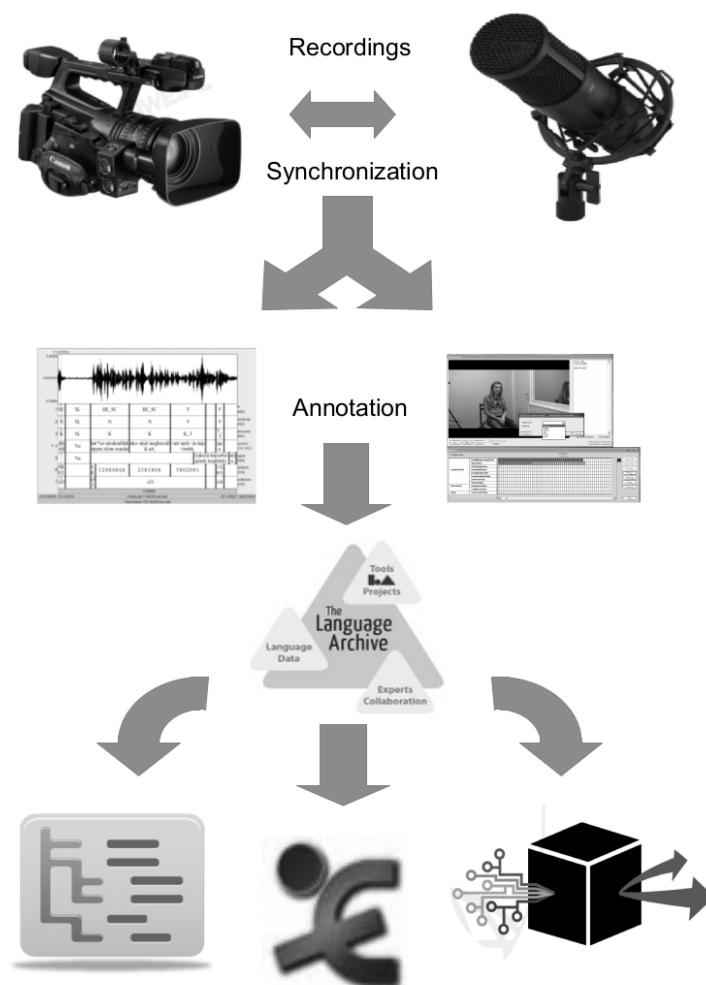


4.2. ábra. A nevetések percenkénti előfordulása a formális és az informális dialógusokban.

Ezek az egymással kontrasztív párokba rendezhető formális és informális interjúfelvételek jelentik a korpusz legértékesebb, legváltozatosabb feldolgozási lehetőséget magukban rejtő részét, amely a jelen disszertáció írásáig is számos, az elemzési szempontok széles spektrumát átfogó tudományos munka alapjául szolgált.

4.1.2. A korpusz feldolgozásának folyamata

A 4.3. ábra a korpusz feldolgozásának lépéseit szemlélteti. A felvételek elkészítése után az első lépést a különböző felvevőeszközök médiaanyagának időben történő összeszinkronizálása jelentette, amely az egyes „jeleneteket” elválasztó, sípoló hangjelzések alapján történő vágásokkal valósult meg. Az annotáció bemenetéül már ez a vágott hang- és képanyag szolgált, amelyet az annotáló programok (Praat és Qannot) igényei szerint konvertáltunk tovább.



4.3. ábra. A korpusz feldolgozásának folyamata.

A annotációs szintek különböző formátumú szöveges kimeneteit első körben a Praat program által használt TextGrid formátumban egységsítettük, majd a videófelvevételek vágásának pontatlanságait kompenzálendő további szinkronizálást végeztünk rajtuk, mielőtt a teljes anyag összefűzésre került. Az összefűzés eredményeként születő TextGrid fájlokat az ELAN program, illetve a korpusznak a The Language Archive felületén elérést biztosító keretrendszer által is támogatott EAF formátumba konvertáltuk tovább. A médiafájlok linkelésénél a felvételek szinkronizált lejátszása az alábbi formában volt paraméterezhető:

```
<HEADER MEDIA_FILE="" TIME_UNITS="milliseconds">
  <MEDIA_DESCRIPTOR MEDIA_URL="file:///path_to_file/audio.wav"
  MIME_TYPE="audio/x-wav"
  RELATIVE_MEDIA_URL="./audio.wav" TIME_ORIGIN="142"/>
  <MEDIA_DESCRIPTOR MEDIA_URL="file:///path_to_file/video1.mp4"
```

```

MIME_TYPE="video/*"
RELATIVE_MEDIA_URL="./video1.mp4" TIME_ORIGIN="0"/>
<MEDIA_DESCRIPTOR MEDIA_URL=file:///path_to_file/video2.mp4"
MIME_TYPE="video/*"
RELATIVE_MEDIA_URL="./video2.mp4" TIME_ORIGIN="82"/>
</HEADER>

```

A metaadatok elkészítésére az ARBIL programot (Withers, 2012) használtuk fel, majd a LAMUS (Broeder et al., 2006) rendszeren keresztül feltöltve a The Language Archive web alapú, az IMDI metaadatsémát (Broeder & Wittenburg, 2006) követő böngészőjében tettünk közzé minden anyagot, ahol az érdeklődő kutatók a TROVA kereső⁹ felületén hajthatnak végre lekérdezéseket, de az egyes annotációk és megfelelő jogosultsággal a médiafájlok is böngészhetők az ANNEX (Berck & Russel, 2006) interfész használatával.

Az EAF formátum további konvertálásával (az ELAN program erre számos beépített lehetőséget biztosít) az adatok alternatív, más kutatási eszközök és az egyszerű lekérdezéseken túlmutató célok (gépi tanulás, rejtett időbeli minták feltérképezése) által támogatott reprezentálása is megvalósítható. Egy a HuComTech kutatócsoport és az Izlandi Egyetemen *Human Behavior Laboratory* nevű intézete¹⁰ között kialakult, a legújabb kutatási irányoknak keretet adó együttműködés keretében például fontossá vált, hogy a korpuszban tárolt címkék a rejtett időbeli minták feltérképezésére szolgáló THEME program (Magnusson, 2000) által is elemezhetővé váljanak.

4.1.3. A korpusz annotációs sémái

A korpusz annotációs sémáinak általános funkciója, hogy egy adott, a kommunikáció szempontjából releváns jelenség különböző aspektusainak leírását véges számú attribútumok (pl. arckifejezés, dallammenet, tekintetirány stb.) és az attribútumok által felvehető értékek (a dallammenet esetén pl.: emelkedő, eső stb.) meghatározása által szabványosítsa. A annotáció folyamán ezek az előre meghatározott attribútumok kapnak egy adott időpillanathoz vagy időintervallumhoz hozzátársított, az elemzés eredményének leginkább megfeleltethető értéket a lehetséges értékek halmazából. A konkrét annotációs sémák nem csak sémában használt attribútumok (az annotáció taxonómiája), de az egyéb metaadatok tekintetében is különbözőek lehetnek. A 4.1. táblázat a vizsgált jelenség (az annotáció tartalma), az eredményül kapott adatok – az elemzés jelle-

⁹<http://tla.mpi.nl/tools/tla-tools/trova>

¹⁰<http://hbl.hi.is/>

géből fakadó – természete (deskriptív vagy interpretáció), az elemzéshez forrásaként használt modalitás (vizuális vagy auditoriális), az elemzés módszere (manuális vagy automatikus) és a felhasznált szoftveres környezet alapján jellemzi a HuComTech korpuszban rendelkezésre álló anyagot.

4.1. táblázat. A HuComTech multimodális korpusz annotációs sémáinak összefoglaló táblázata.

Tartalom	Elemzés	Forrás	Módszer	Eszköz
szöveges átiratok	deskriptív	hang	manuális	Praat
non-verbális gesztusok	deskriptív	videó	manuális	Qannot
prozódia	deskriptív	hang	automatikus	Praat
szintaxis	deskriptív	szöveges átirat	manuális	Praat
megfigyelt érzelmek	interpretáció	hang	manuális	Praat
		videó		Qannot
		hang + videó		
pragmatika	interpretáció	videó	manuális	Qannot
		hang + videó		

A táblázatból látható, hogy a prozódiai annotáció kivételével minden más annotációtípus manuális módszerekkel, az elemzésre és az elkészült elemzések ellenőrzésére kijelölt személyek (továbbiakban: annotátorok) megfigyelései által valósult meg. A Praat program annotációs felületén készült manuális annotációk (szöveges átiratok, szintaxis, érzelmek) esetében az ellenőrzés automatikus eszközökkel is kiegészült, ami utólagos korrekciókkal kompenzálta, hogy az attribútumok által felvehető értékek korlátozására az annotáció folyamán technikailag nem volt lehetőség.¹¹ A prozódia – a Praat program beépített szkript nyelvén implementált – automatikus annotációjával kapcsolatban a bevezetésben már feltártam néhány részletet. Mivel a módszer kidolgozása a disszertáció közvetlen céljai közé tartozik, bővebb kifejtésére a korpusz általános bemutatása után, a 4.2. részben kerül majd sor.

A létrejött adatok természetével kapcsolatban fontos megkülönböztetés, hogy (a) egy fizikai vagy nyelvészeti szempontból egzaktan megragadható jelenség deskriptív leírását vagy pedig (b) valamilyen, az utóbbi jelenségek interpretációján alapuló megfigyelést közvetítenek. Az első csoport prototipikus esetét jelentik a non-verbális gesztusok (pl. a tekintet iránya, a szemöldök, a kéz, a fej mozgása stb.) és a szöveges átiratok (a társalgás során elhangzott szavak

¹¹ A videók annotálására szolgáló, az annotációs sémát egy külső XML állományból beolvasó Qannot program már eleve csak felvehető értékek tartományából engedett választani.

sorozata), amelyek kellően egzaktak és jól megfigyelhetők ahhoz, hogy az esemény leírásakor csak bizonyos szélsőséges (vagy köztes) esetekben támaszkodnak a megfigyelő (vitatható) értelmezésére. A szintaxis és a prozódia esetében már kevésbé egyértelmű ugyanez a kérdés, de az adatok az ő esetükben sem lépik túl a leírás szintjét. A *leíró* nyelvtan szabályait követő szintaktikai annotáció nagyon is egzakt és gépiesen alkalmazható szabályok segítségével tárja fel a szöveges átiratokban transzkribált megnyilatkozások mondszerkezetét, a prozódiai annotáció pedig az F0 és intenzitásadatok hasonlóan deskriptív leírására szorítkozik.¹² A pragmatikai annotációban jelölt beszédaktusok, illetve az érzelmi címkéket tartalmazó annotációk azonban már a közvetlenül megfigyelhető jelenségek értelmezésének dimenziójába vezetnek át, ahol az interpretáció alapját tulajdonképpen a kommunikációs események előző csoportba sorolt leírásai jelentik.¹³

A elemzés forrásaként használt modalitások kapcsán külön kiemelném a pragmatikai annotáció „unimodális” verzióját, amelyben az annotátoroknak kizárólag a képanyag (egy hang nélkül lejátszott videó) alapján, tehát a verbális információkhoz való hozzáférés nélkül kellett a felvételeket pragmatikai szempontból annotálniuk.¹⁴ A disszertáció keretében végrehajtott percepciók kísérletek hasonlóan unimodális kondíciók mellett: a verbális és a vizuális információk kizárásával történtek¹⁵, amit az első kísérletekben Baiat & Szekrényes (2012) esettanulmánya nyomán a HuComTech korpusz informális dialógusainak mintáját követő, de perzsa nyelven folyó dialógusok és a perzsa nyelvet nem ismerő kísérleti alanyok kombinációja biztosított. A kísérletek során használt egyéb módszerekkel és lebonyolításuk részleteivel az 5. fejezet foglalkozik majd bővebben.

4.2. A prozódia annotációjával kapcsolatos fejlesztések

Az annotáció fejlesztésekor fontos szempontot jelentett a beszéd prozódiai jelenségeinek azon – a percepció számára leginkább releváns – természete, hogy nem konkrét időpillanatokhoz vagy bizonyos szegmentális egységekhez köthető értékek formájában (mint amilyen egy adott

¹² A leírásra szorítkozás nem összekeverendő az „elmélet-semlegesség” korábban felvetett igényével, amelyre egy szigorú értelmezésben már a *mondat* vagy a *hangmagasság* fogalmának alkalmazásakor sem vagyunk maradéktalanul tekintettel.

¹³ Például a megnyilatkozás verbális tartalma és prozódiaja, továbbá (a kommunikáció multimodális természetére tekintettel) a non-verbális gesztusuk alapján határozzuk meg beszédaktus illokúcióját, de ugyanezek alapján a beszélő érzelmi állapotaira is következtethetünk.

¹⁴ A táblázatból látszik, hogy hasonló módon történt az érzelmek annotációja is: csak kép alapján, csak hang alapján, illetve a két modalitást együtt használva kellett a beszélő érzelmi állapotát interpretálni.

¹⁵ A teljes, 50 órás anyagon véghezvitt, a vizualitás helyett a prozódiaán alapuló annotáció természetesen nem szerepel a kutatás céljai között.

t időpontban érzékelt, Hertzben kifejezhető hangmagasság), hanem a pillanatnyi értékek változását, tendenciáit tartalmazó időbeli, szupraszegmentális folyamatként ragadhatóak meg. A változás folyamatának elsőbbsége az egyes fizikai jellemzők konkrét, időhöz kötött értékeivel szemben leginkább a *változás* hiányakor, a prozódiai jellemzők szélsőségesen „lapos”, monoton megvalósulásakor érzékelhető, például egy olyan géppel szintetizált beszéd esetén, amelyben az érintett jellemzők állandóak, nincs moduláció, az eredményezett hang prozódiával tehát effektíve nem is rendelkezik.¹⁶ A HuComTech korpusz annotációs címkéi az interakciók prozódiai szerkezetének elemeit a „mozgásnak” ebben a relatív viszonyrendszerében igyekeznek reprezentálni: a moduláció iránya és mértéke szerint jellemezve a vizsgálat tárgyaként kijelölt jellemzőket (dallam, intenzitás és tempó) állandó tekintettel a résztvevők beszédmódjának egyedi sajátosságaira. A multimodális kommunikáció non-verbális modalitásai között fontos szerepet betöltő prozódia elemzése és reprezentációja nem csupán a disszertáció céljai szempontjából volt elengedhetetlen, de az itt bemutatásra kerülő módszerek kivitelezése által a HuComTech-korpusz általános felhasználási lehetőségei is jelentős mértékben gazdagabbá váltak.

A módszer technikai kialakításánál fontos szempontot jelentett az univerzális hasznosíthatóság is, amelynek eredménye, hogy az annotációt végző algoritmus tetszőleges, a minimális formai kritériumoknak (lásd 4.2.1.3. szakasz) eleget tevő korpusz prozódiai elemzéséhez felhasználható. A HuComTech korpusz hangfelvételein való tesztelés után a SegCor projekt¹⁷ közreműködésével, a FOLK korpusz (Schmidt, 2016) német nyelvű, változatos kondíciók között (2–14 adatközlővel) készített hangfelvételeinek elemzése során további javítások és átdolgozások valósultak meg, amelyeknek a program jelenlegi flexibilitását köszönheti.

Az automatikus prozódiai annotációval kapcsolatos fejlesztések kezdeti szakaszának rész-eredményei a HuComTech korpusz feldolgozásának új lehetőségeként a VIII. Magyar Számítógépes Nyelvészeti Konferencián kaptak először nyilvánosságot (Szekrényes et al., 2011). A későbbi publikációk elsősorban a beszéddallam automatikus lejegyzésére szolgáló eljárás fejlesztésének teoretikus és módszertani hátterét (Szekrényes, 2014), illetve a konkrét technikai kivitelezés részleteit (Szekrényes, 2015) tárgyalják. A Praat beszédfeldolgozó program szkript nyelvén implementált, eredetileg ProsoTool névre keresztelt alkalmazás (a továbbiakban tehát: ProsoTool) aktuális verziója a beszéd szupraszegmentális jelenségei közül a beszéddallam

¹⁶ Alapfrekvenciája és intenzitása természetesen minden hanghullámnak van, de dallamról, tempóról vagy ritmusról az értékek változásának hiányában aligha beszélhetünk, hacsak az állandóságot, a tempónélküliséget nem tekintjük a „kottában” afféle tudatosan választott null-értéknek. Egy érdekes, a „gépi hangnál” is szélsőségesebb párhuzamként említhető a kortárs experimentális zene világából John Cage 4:33 című darabja, amely egyetlen prozódia sajátosságként csak az időtartamot őrizte meg, mivel a zenészek az előadás során hozzá sem értek a hangszereikhez.

¹⁷ <http://www1.ids-mannheim.de/prag/muendlichekorpora/segcor.html>

elemzését és feliratozását végzi el automatikusan, de történtek fejlesztések a beszéd intenzitásának és tempójának elemzésére is. Később az e-magyar projekt¹⁸ (Várad et al., 2017) keretében a beszélődiarizálás kimenetére épülő modulként lett elérhető (Kornai & Szekrényes, 2017) a terület kutatói számára emPros név alatt. A prozódiai annotáció módszereire támaszkodó tudományos eredmények három könyvfejezet (Hunyadi et al., 2016a,b; Kovács & Szekrényes, 2019), két konferencia közlemény (Kovács & Várad, 2017; Szekrényes & Kovács, 2017) és egy sikeresen megvédett doktori disszertáció (Abuczki, 2014) keretében kerültek publikálásra.

4.2.1. A beszéddallam annotációja

Az beszéddallam annotációjára kidolgozott modul fejlesztését technikai szempontból leginkább Christophe d'Alessandro és Piet Mertens a tonális kontúrok pszicho-akusztikai alapokon (Hart, 1976) történő stilizálására kidolgozott (d'Alessandro & Mertens, 1994), később Prosogram néven (Mertens, 2004) implementált módszere inspirálta, de az alaphang modulációinak kategorizálására használt módszereket tekintve sokat merített a Tilt¹⁹ modell (Taylor, 2000) paramétereiből is. Az előbbi eljárás lényege, hogy a dallam ténylegesen érzékelhető modulációit a beszéd intenzitása és bizonyos – a zenetudományból is kölcsönzött – küszöbértékek segítségével emeli ki a frekvenciaértékek csak géppel mérhető változásai közül. Az eredményül kapott reprezentáció az alaphang szótag-szinten stilizált mozgásait (az eredeti adatsorok értékpárokra redukált vektorait) tartalmazza. Saját algoritmusom kialakítása során a szótagok szintjén túllépve a beszéddallam holisztikusabb, adott esetben akár több szótagot is egyetlen „dallam-eseményben” integráló leképezésére törekedtem, ahol az eredményül kapott annotációs címkék a hanglejtés formájáról (emelkedő, eső, ereszkedő stb.) és a beszélő egyedi hangterjedelmében elhelyezett relatív (kezdő és záró) hangmagasságáról adnak számot. A dallamformák kategorizálása során figyelembe vett küszöbértékek számításához a beszélők egyedi hangterjedelmét, átlagos hangmagasságát és hangmagasság-ingadozását vettem alapul, de fontos támpontot jelentettek a percepciók kísérletek eredményei is.

Mivel a beszéddallam feldolgozásáért felelős algoritmus az intenzitás és sok tekintetben a beszédtempó annotációjához is mintául szolgált²⁰, célszerűnek találtam az annotációs eljárás általános működését a beszéddallamra vonatkozó konkrét példákon keresztül bemutatni. Az intenzitásról és a beszédtempóról szóló alfejezetek (sorrendben 4.2.2. és 4.2.3.) pedig inkább az

¹⁸<http://e-magyar.hu/hu>

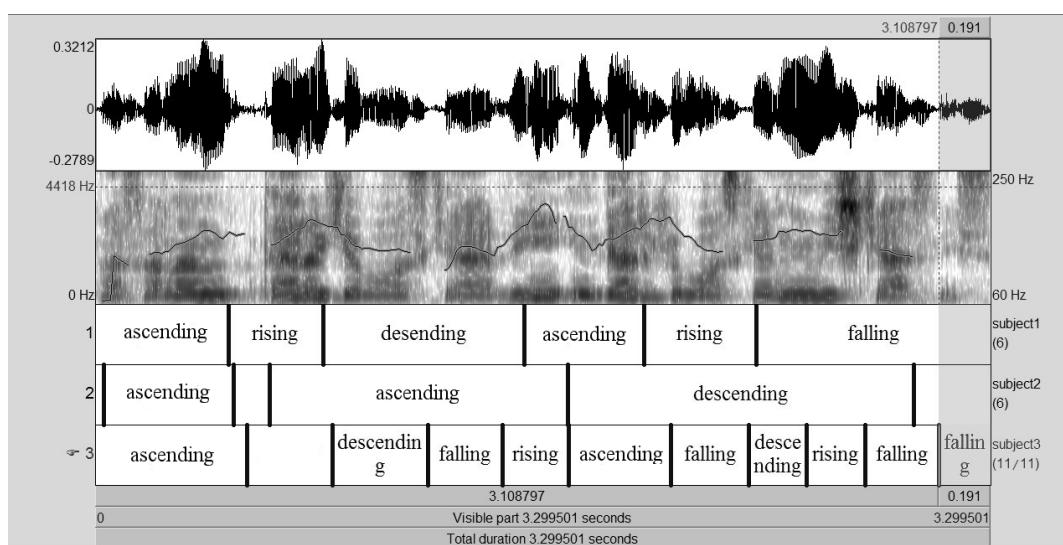
¹⁹http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/c16909.htm

²⁰ A fejlesztés időrendjében is megelőzte az utóbbiakat.

eltérésekre, az utóbbi két jellemző feldolgozásához szükséges egyedi megoldásokra fókuszálnak.

4.2.1.1. Manuális annotáció (kísérletek)

A prozódia percepciójára vonatkozó elméleti bevezetőkhez és az automatikus annotáció céljaihoz kiváló illusztrációként szolgálnak az alkalmazás fejlesztése során végzett, a prozódia manuális annotációjával kapcsolatos kísérletezés eredményei. Ezek a próbálkozások nem az 5.2. fejezetben bemutatásra kerülő percepció tesztek paradigmáját követték, mivel az alanyoknak – akik nyelvészeti háttérrel rendelkező, képzett annotátorok voltak – egy előre kidolgozott, később az automatikus annotációnál is felhasznált séma szerint kellett elemezniük egy néhány másodperc hosszúságú hangfelvételt. Az elemzést nem az alapfrekvencia adatok, hanem hallás alapján végezték: a dallammenet iránya és a moduláció mértéke szerint szegmentálták és felcímkézték a felvételen hallható beszédrészletet. Az annotációs sémában az alábbi, a dallammenet formájára vonatkozó kategóriák álltak rendelkezésükre: *rising* (gyors/jelentős emelkedés), *falling* (gyors/jelentős esés), *ascending* (lassú/kisebb mértékű emelkedés), *descending* (lassú/kisebb mértékű ereszkedés), *stagnant* (nincs változás: monoton dallammenet).



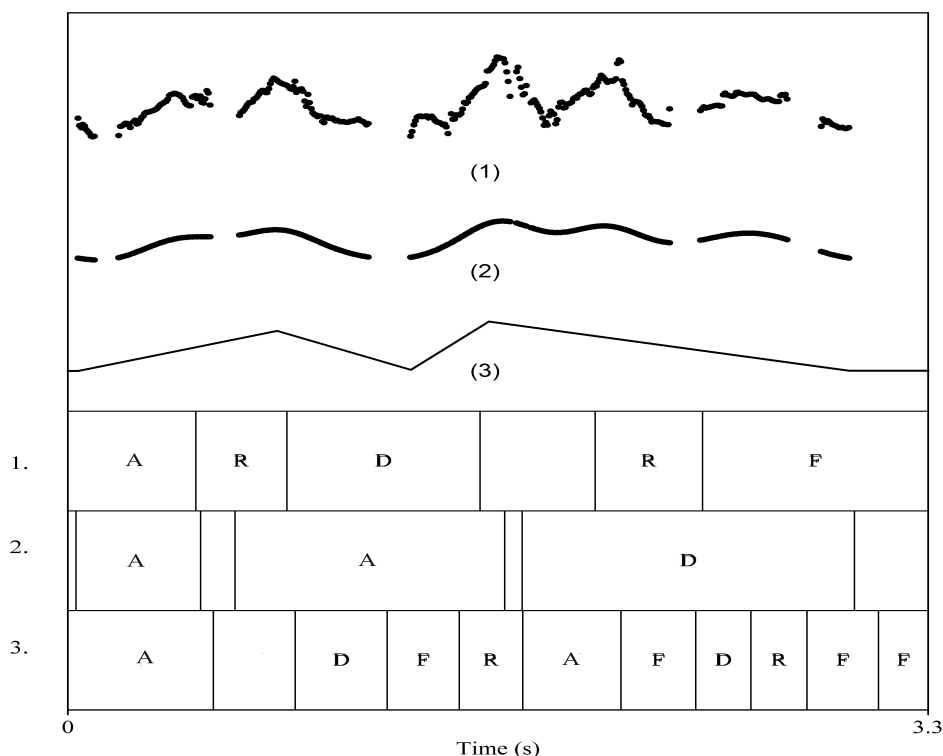
4.4. ábra. A beszéddallam manuális annotációjának eredményei.

A 4.4. ábra alsó felében három annotátor (fentről lefelé: *subject1*, *subject2*, *subject3*) a hangfelvétel idővonalát követő címkézésének eredményei láthatók egymással párhuzamosan. Az ábra felső része a felvétel hullámformáját (fent) és hangszínképét (lent), a hangszínképre rajzolt görbe pedig a Praat program autokorrelációs módszerével (*minimum Pitch*: 75 Hz, *Maximum Pitch*: 600 Hz alapértelmezett beállítások mellett) mért alapfrekvencia-értékeit szemlélteti.

Az eredménye elemzését kétféle szempontból: a gépi mérések és az annotátorok megfigyelései, illetve az utóbbiak egymás között való eltérései alapján végezhetjük el. Az egyes címkékre a különböző megfigyelőkhöz tartozó annotációs szintek (vertikális tengely) számából és a címke üres-intervallumokat ignoráló sorszámából (horizontális tengely) kombinált koordinátákkal fogok hivatkozni.²¹ A felvétel legelején (az [1,1], [2,1], [3,1] szegmentumokban) az annotátorok értékelése szinte tökéletes szinkronban van egymással és az alaphang adatokkal, habár az F0-értékek néhány Hertzes „bukdácsolását” a percepció természetesen már itt is ignorálja. A soron következő dallammenetek szegmentálásában és értékelésében viszont jelentősebb különbségek mutatkoznak. A második annotátor például a dallammenet kisebb változásait a többiekhez képest hosszabb, tehát kisebb felbontású trendekben integrálja. A felvétel prozódiai szerkezetét három fő dallammenetre osztja fel, amiből az utolsó kettő ([2,2] és [2,3]) formáját az alaphang a többi annotátor által is megfigyelt ([1,4] és [3,5]) jelentősebb emelkedése – ami egyben az alaphang egész felvételre értelmezhető maximuma is – határozza meg. Felosztásának eredménye: az emelkedés csúcspontjáig *ascending* [2,2] majd a csúcstól lefelé *descending* [2,3], ahol az emelkedés magába integrálja (mintegy „beolvasztja”) az alaphang adatok alapján is megfigyelhető csökkenését és megszakadását, amit az első annotátor *descending* [1,3], a legnagyobb felbontást produkáló harmadik annotátor pedig *descending* és *falling* ([3,3] és [3,4]) címkékkel írt le. A később jelölt ereszkedés [2,3] hasonlóképp figyelmen kívül hagyja az alaphang többiek által észlelt ([1,4], [1,5], [3,6], [3,9]) ideiglenes emelkedéseit.

Ha az alaphang frekvencia görbe 4.5. ábrán látható (1) eredeti formájának (2) simítását, majd a simított görbe (3) stilizálását a Praat program segítségével elvégezzünk, akkor az adatok is jobban kidomborítják az alaphang frekvencia-értékek változásának azt a tendenciáját, amit a második annotátor az alaphang ereszkedéseként [2,3] ragadott meg. Mindez nem jelenti azt, hogy a többiek (nagyobb felbontású) észlelései nem relevánsak, és hogy a második annotátor percepcióját követve nem veszítünk információkat. Ugyanakkor az alaphang változásának bizonyos tendenciái, amit az eredeti részletek rejtve hagynak, csak az alaphang kisebb felbontású feldolgozása után világíthatók meg. Az értékelés összegzéseként tehát megállapíthatjuk, hogy az elemzés eredményéhez nem kizárólag a vizsgált jelenség attribútumai, de a megfigyelő „pozíciója”, a választott megfigyelési perspektíva is jelentékenyen hozzájárul.

²¹ Az [1,3] például a *subject1* által jelölt *descending* címkéjű szegmentumra vonatkozik.



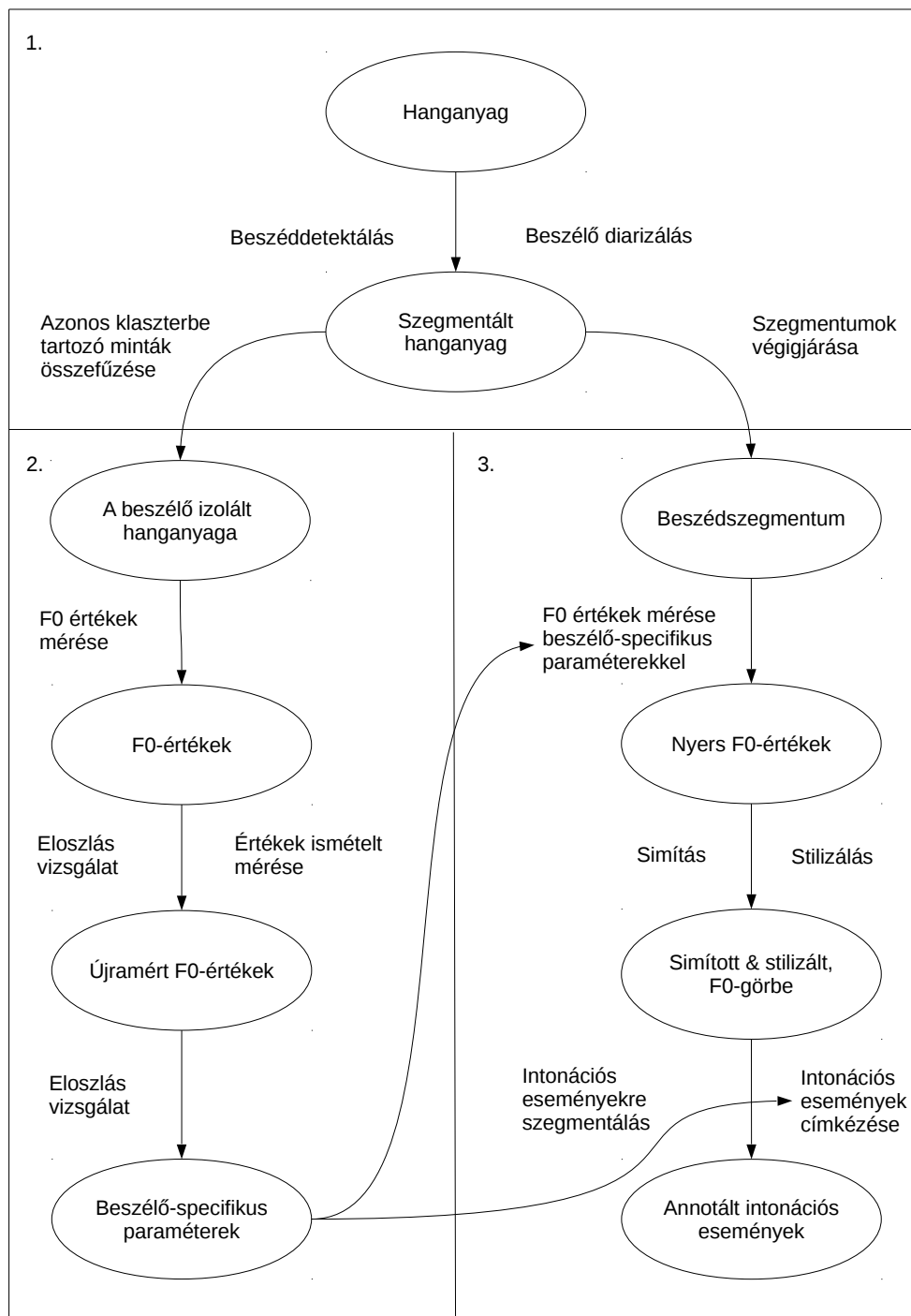
4.5. ábra. Az adatok és a manuális annotációk összevetése az F0-görbe simítása és stilizálása után. A manuális annotációk címkéit (az ábra alsó részén, a görbékkel párhuzamos időtengelyeken látható szimbólumok) ezúttal a kategóriák kezdőbetűi helyettesítik: (F)all, (R)ise, (A)scending, (D)escending.

Az elemzés kivitelezése és az eredményül kapott adatok felhasználhatóságának szempontjából a manuális annotáció problematikusságát nem csupán a feladat nehézsége²² és a megfigyelő személyek szubjektív ítéletei között különbség jelenti, de az ugyanazon személytől származó megfigyelések belső, a dallammenet leírásának részletességében (felbontásában) megmutatkozó következtelenség lehetősége is. A szubjektív érzékítéletek divergens természetét tekintve természetesen az automatikus prozódiai annotáció eredményeitől sem várható, hogy tetszőleges megfigyelő személy benyomásaival fedésbe hozhatók legyenek. Fontos szempont volt viszont az adatok alapján létrehozott reprezentáció belső konzisztenciája, a szubjektív benyomásokkal való összevethetősége és ellenőrizhetősége. Mivel egy teljesen automatikus, a mért adatokra támaszkodó módszerről beszélünk, az annotációs címkék és az azonos jelenségre vonatkozó érzékítéletek között előforduló diszkrepanciák bizonyos esetekben a közvetlen mérések során bekövetkező hibáknak is tulajdoníthatók, amelyeket nem mindig lehetett a méréseket feldolgozó eljárásokban kiküszöbölni.

²² Az annotátorok számára ez a rövid, 3 másodperc hosszú felvétel – hallás alapján történő – elemzése is komoly megterhelést jelentett.

4.2.1.2. Az algoritmus rövid leírása

A disszertáció keretében kidolgozott, a beszéddallam automatikus annotációját végző algoritmus működése az alábbi, a 4.6. ábrán is szemléltetésre kerülő lépésekben foglalható össze (zárójelben jelzem, hogy melyik művelet történik automatikusan és melyik igényel jelenleg manuális előkészítést):



4.6. ábra. A ProsoTool algoritmusában felhasznált eljárások három blokkra (1. előfeldolgozás, 2. beszélő-specifikus paraméterek meghatározása, 3. szegmentumok feldolgozása) felosztható kapcsolat-rendszere.

1. Előfeldolgozás:

- a) A beszédaktivitás detektálása (manuális)
- b) A beszélők diarizálása (manuális)
- c) Az egyes beszélőkhöz tartozó szegmensek külön hangfájlokban való izolálása az előző lépések (1/a és 1/b) kimenete alapján (automatikus)

2. Beszélő-specifikus paraméterek meghatározása:

- a) Az F0-értékek kinyerése az adott beszélőhöz tartozó összes mintából az 1/c lépés kimenetének felhasználásával (automatikus)
- b) Az F0-értékek eloszlásának vizsgálata és az értékek újramérése az aktuális beszélő hangterjedelmének figyelembevételével (automatikus)
- c) Az F0-értékek eloszlásának ismételt vizsgálata, küszöbértékek számítása az előző (2/b) lépésben kapott értékek alapján. (automatikus)

3. Szegmentumok feldolgozása:

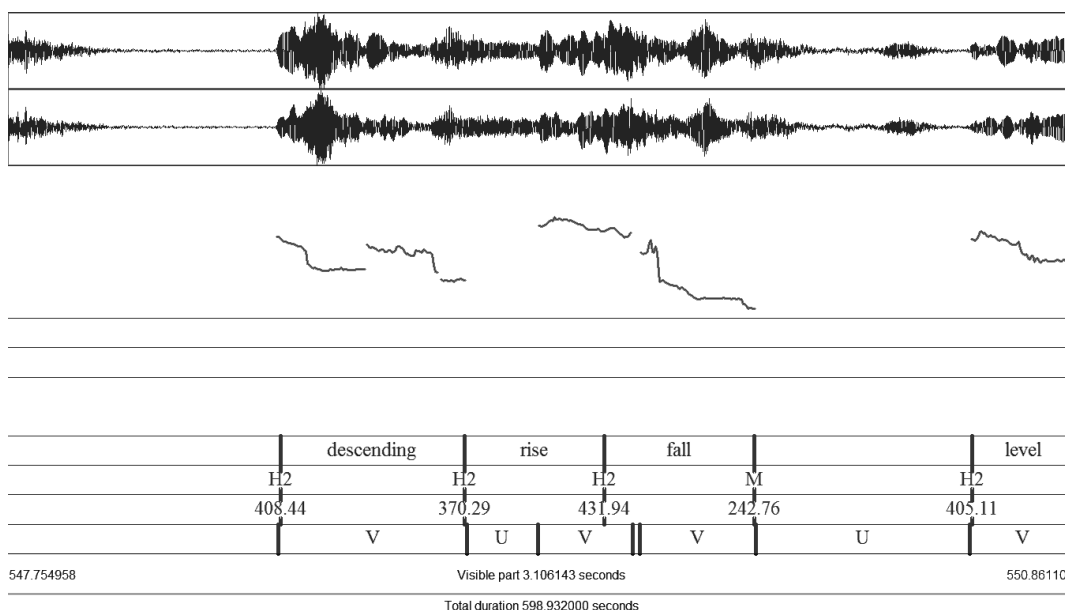
- a) F0-értékek szegmentumonkénti kinyerése (automatikus)
- b) Az F0-görbék simítása (automatikus)
- c) A simított F0-görbék stilizálása (automatikus)
- d) A dallamformák szegmentálása a stilizált F0-görbék alapján (automatikus)
- e) A létrejött dallam-szegmentumok osztályozása és címkézése az előző (3/d) lépés és a 2/c lépés eredményei alapján (automatikus)

A szkript bemenete egy a társalgást rögzítő egy csatornás (mono), WAV formátumú hangfájl és egy hozzá tartozó annotáció, amely a felvételen elhangzó megnyilatkozások időbeli pozícióját beszélőnként külön tengelyen (annotációs szinten) tartalmazza. Az ehhez szükséges előfeldolgozási feladatok (beszéddetektálás és beszélő diarizálás) tehát külső feldolgozó programok használatát vagy manuális munkát igényelnek, de a beszélők számára, a megnyilatkozások tagolására és feliratozására vonatkozóan nincs semmilyen megkötés. Az üres címkék a kérdéses beszélő hallgatásaként értelmeződnek. Az annotáció elkészítéséhez vagy a meglévők konvertálásához a következő, 4.2.1.3. szakaszban található bővebb információ. Az annotáció hiányában a szkript a hangfájl teljes hangzó tartalmát egyetlen beszélő klaszterhez társítja a beszédszünetek automatikus detektálása után, amelyhez a Praat program "To TextGrid (silences)", a 3. fejezetben már bemutatott függvényét használja fel. Az szkript működését az alábbi, kötelező argumentumként beállítandó paraméterekkel szabályozhatjuk:

- INPUT_PATH: A bemeneti könyvtár elérési útja. Értéktartomány: sztring
- STYLIZATION: Az F0-görbe stilizálásának felbontása. Magasabb érték erősebb stilizálást (durvább felbontást) eredményez. Értéktartomány: egész szám. Optimális beállítás: 2.
- SMOOTHING: Az F0-görbe simítását meghatározó paraméter. Alacsonyabb érték erősebb simítást eredményez. Értéktartomány: valós szám. Optimális beállítás: 1.5.
- PITCH_EXTRACTION: Az F0-mérések paramétereinek (alsó és felső küszöb) meghatározási módja. Értéktartomány: *standard* vagy *dynamic*
- OPERATING_SYSTEM: Futtatási környezet. Megadása könyvtárműveletek miatt szükséges. Értéktartomány: *windows* vagy *unix*

Szintén az előfeldolgozás részét képezi az azonos beszélőkhöz tartozó szegmentumok külön hangfájlban való összefűzése (a beszélők hangjának izolálása), aminek eredménye a második blokkba tartozó lépések bemenetét adják a beszélő-specifikus paraméterek meghatározásához (lásd 4.2.1.4. szakasz). Ezeket a paramétereket a harmadik blokkban fogjuk hasznosítani az egyes megnyilatkozásokat képező beszédsegmentumok elemzéséhez: először a nyers F0-értékek mérésénél, majd a görbék simítása, stilizálása és szegmentálása után az intonációs események osztályzásánál (lásd 4.2.1.6. szakasz).

A feldolgozás után kapott Praat TextGrid formátumú kimenetet egy a FOLK korpuszból (Schmidt, 2016) vett példa alapján a 4.7. ábra szemlélteti, ahol az automatikus annotáció eredménye négy párhuzamos idővonalra felosztva található meg. Fentről lefelé haladva:



4.7. ábra. Az F0 annotációjának kimenete a Praat program szerkesztő felületén.

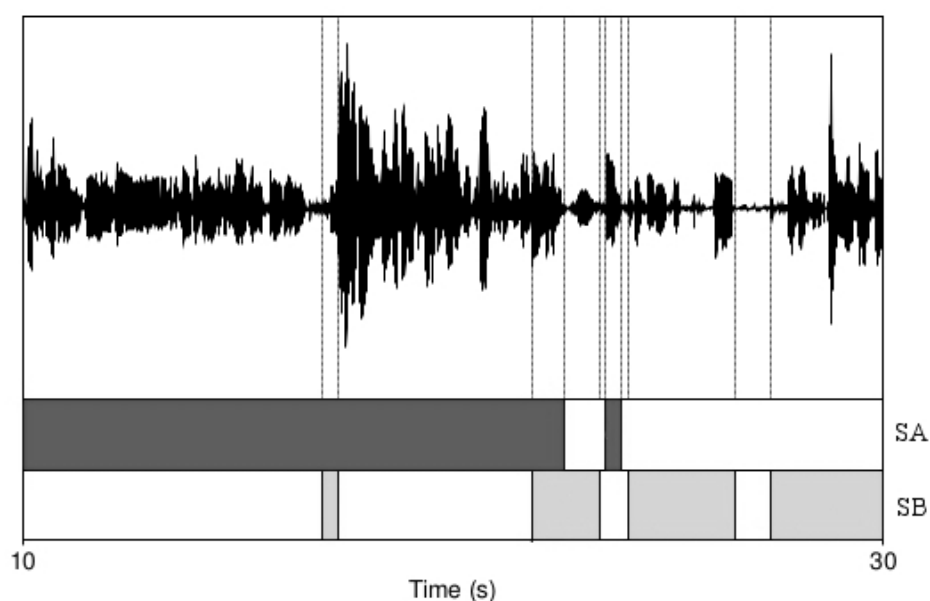
- A dallam modulációjának egy adott szegmentumra vonatkozó formája, amely az alábbi értékek valamelyike lehet: *rising* (gyors/jelentős emelkedés), *falling* (gyors/jelentős esés), *ascending* (lassú/kisebb mértékű emelkedés), *descending* (lassú/kisebb mértékű ereszkedés), *level* (nincs jelentősebb moduláció: szinttartó dallammenet).
- Ugyanezen szegmentum kezdetén és végén mért relatív hangmagasság a beszélő egyedi hangterjedelmének a frekvencia értékek eloszlása alapján 5 részre felosztott skáláján. Lehetséges értékek a mélyebbtől a magasabb tartományok irányában: L_2 (legmélyebb), L_1 (mély), M (közép), H_1 (magas), H_2 (legmagasabb).
- A fentebb leírt hangmagasságok Hertzben megadott eredeti értéke.
- A megnyilatkozás zöngés („V”) és zöngétlen („U”) szakaszainak határai és időtartama.

4.2.1.3. Előfeldolgozás és bemeneti kritériumok

A ProsoTool eredetileg dialógusok prozódiai szerkezetének automatikus lejegyzésére lett optimalizálva, de a FOLK korpuszon végzett tesztek után a beszélők izolálását végző eljárás jelentős átalakításra került, így – az annotációs szintek dinamikus feldolgozása és az átfedő beszédek rekurzív kezelése révén – kellően flexibilissé vált ahhoz, hogy tetszőleges számú beszélőre is működőképes legyen. Az eljárás lényege, hogy az interakcióban résztvevő beszélők hangját izoláljuk a többi beszélőhöz, az átfedő beszédekhez és a beszédsszünetekhez tartozó szegmensektől. A későbbi eljárások eredményeit jelentősen torzítaná ugyanis, ha a beszélőnként változó hangterjedelem jellemzésére a többi beszélőre vonatkozó mérések és a beszédsszünetek időtartama is befolyást gyakorolna.

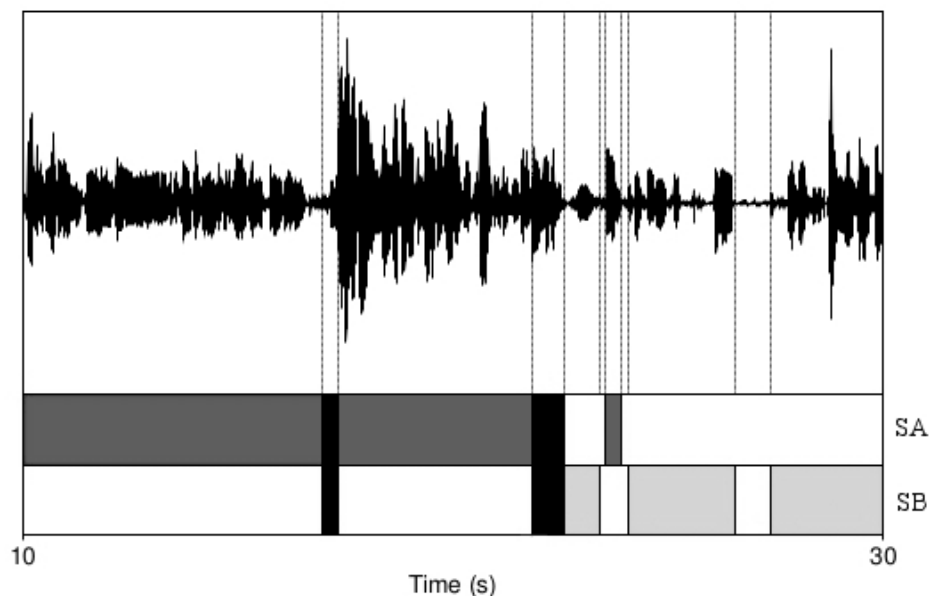
A szünetek és a beszélőváltás folyamatának annotálása történhet manuálisan, vagy a 3. fejezetben bemutatásra került, a beszédaktivitás detektálását és a beszélők diarizálását végző eszközökkel. A legtöbb beszélt nyelvi korpusz esetében – így a fejlesztés elsődleges tesztanyagát jelentő HuComTech-korpuszban is – a szükséges információt tartalmazó annotációk, ha nem is mindig explicit formában, de már eleve rendelkezésre állnak. „Explicit forma” alatt itt a beszélőváltás akusztikai szempontból történő reprezentációját (lásd. 4.8. ábra) kell érteni, ahol a szegmentumok lehetséges határai a beszélőváltások vagy a beszédsszünetek helyére korlátozódnak. Könnyen előfordulhat, hogy a vizsgált élőnyelvi korpuszban rendelkezésre álló annotációk egyike sem felel meg ennek a követelménynek, mivel a beszélők megnyilatkozását más szempontból, például azok verbális tartalma alapján tagolják szintaktikai (mondatok, tagmondatok), lexikai (szavak) vagy még kisebb fonetikai (beszédhangok) egységekre a nyelvészeti szempontból történő elemzés és a szöveges átiratok elkészítésének megkönnyítése érdekében.

A ProsoTool előfeldolgozó algoritmus ezt a problémát az eredeti annotáció szegmentumainak szünettől szünetig, szünettől beszélőváltásig (vagy fordítva), illetve beszélőváltástól beszélőváltásig tartó egyesítése útján küszöböli ki. Az egyetlen bemeneti kritérium tehát, hogy a beszélők megnyilatkozásainak helye tetszőleges felbontásban, de az egymástól (a másik beszélőtől) való elkülöníthetőség érdekében különböző annotációs szinteken jelölve legyen.



4.8. ábra. A beszélőváltás akusztikai szerkezetének sematikus ábrázolása a beszélők („SA” és „SB”) megnyilatkozásainak manuális annotációja alapján.

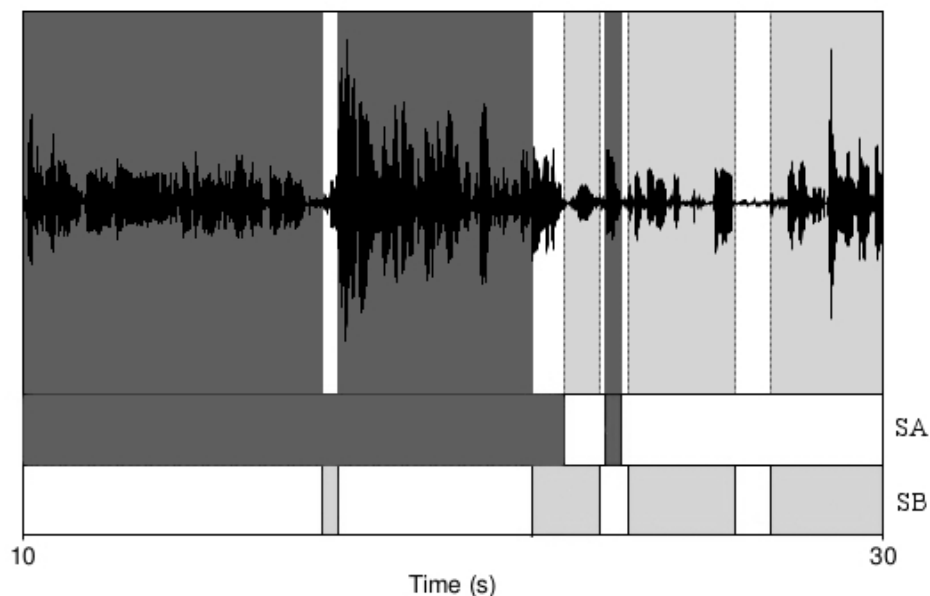
A 4.8. ábra az előfeldolgozás egy köztes, a bemenetként megadott annotációk szükséges átalakításával nyert eredményét szemlélteti. A beszélőváltás folyamata a felvétel hullámformája alatt látható: egymással párhuzamos idővonalakon (sötét és világos szürke sávokkal) jelölve a két beszélő megszólalásainak szünetekkel is tagolt váltakozását és időtartamát. A beszélők hangjának teljes izolálása érdekében az átfedő beszédek kiszűrése jelenti az előfeldolgozás következő lépését, amely az annotációk alapján automatikusan is kivitelezhető, így ezt a műveletet a bemenetként kapott annotáción már a ProsoTool fogja elvégezni helyettünk. A művelet eredménye a 4.9. ábrán látható, ahol a 3. és a 4. annotációs szint a beszélők (1-2. szint) megszólalásainak már csak az átfedő beszédektől (az ábrán feketével jelölt sávok) „megtisztított”, kizárólag az adott beszélő hangját reprezentáló szegmentumokat tartalmazza. A prozódiai elemzés később ugyanezen szegmentumokra korlátozódik, de a ProsoTool aktuális verziójában az átfedő beszédek afféle külön „beszélőként” szintén elemzésre kerülnek.



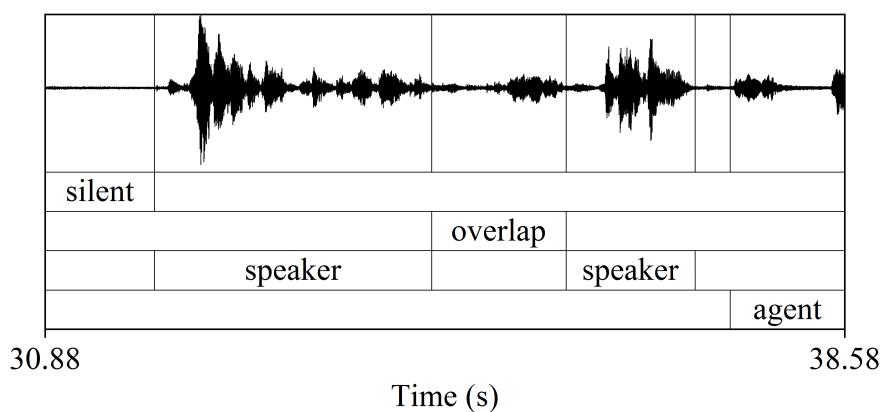
4.9. ábra. Az átfedő beszédek automatikus kiszűrése a beszélőváltás annotációja (4.8. ábra) alapján.

Az előfeldolgozás utolsó, szintén a ProsoTool által automatikusan vezérelt lépése, hogy a beszélők izolált hangját tartalmazó szegmentumokat a 4.10. ábrán látható módon a hangfelvételtől kivágja és egy új hangfájlban fűzi össze a további elemzésekhez. Mint már említettük, ezt a lépést azt teszi szükségessé, hogy a további, a beszélő-specifikus paraméterek meghatározásához szükséges mérések eredményét a többi beszélőhöz tartozó szegmentum, a szünetek és az átfedő beszédek ne torzítsák el.

A további műveletek bemeneteként tehát összesen háromféle állományt készítettünk elő: az eredeti hangfelvételt (valamilyen a Praat által támogatott formátumban), egy ehhez hozzátartozó és azonos elnevezésű, a beszélők izolált hangjának helyét és a szüneteket jelölő annotációt (Praat TextGrid formátumban), végül az egyes beszélők izolált hangját tartalmazó hanganyagot (az eredetivel azonos formátumban). Mivel a prozódia automatikus annotációjának lefuttatása változtatható és eltérő eredményeket produkáló paraméterekkel (ezeket a paramétereket a 4.2.1.2. szakaszban mutattuk be) akár több példányban is indokolt lehet, az elemzés meggyorsítása érdekében, amelyet az előkészítés fentebb bemutatott lépéseinek ismételt végrehajtása jelentősen megnyújtana, az algoritmusnak ezen részét a további műveletektől különválasztottam. Az önálló modulként való alkalmazás másik oka, hogy az előfeldolgozás 4.11. ábrán látható, az átfedő beszédek pontos helyét is tartalmazó kimenete akár önmagában, az alapprozódia elemzésén túlmutató célokra is felhasználható. A disszertáció 6. fejezetében erre konkrét példákat is láthatunk majd.



4.10. ábra. A beszélők hangjának izolálása.



4.11. ábra. A beszélőváltás reprezentációja a ProsoTool kimenetében.

4.2.1.4. A beszélő egyedi hangterjedelmére vonatkozó küszöbértékek számítása

A különböző dallamformák későbbi osztályozása a beszélők izolált hangjának elemzése alapján, dinamikusan meghatározott küszöbértékek segítségével történik. Az alábbi szakasz az F0-értékek számításához és az egyéni hangterjedelem feltérképezéséhez használt módszereket mutatja be.

Az elemzés első lépése, hogy az előfeldolgozás eredményeként kapott, csak az adott beszélő hangját tartalmazó hangfelvételtől az F0-értékek szekvenciáját kimérjük. Ezek a mérések minden esetben (a későbbi műveletekben is) a Praat programban implementált autokorrelációs

módszerrel (Boersma, 1993) történnek. A Praat ide vonatkozó („To Pitch...” kezdetű) parancsainak kimenete egy speciális objektumtípus (*Pitch*), ami a hullámforma egy adott (az aktuális elemzési ablak által lefedett) szegmenséhez társítható lehetséges alapfrekvencia-értékeket (*candidates*), illetve ezeknek az értékek valószínűségét (*strength*) tartalmazza:

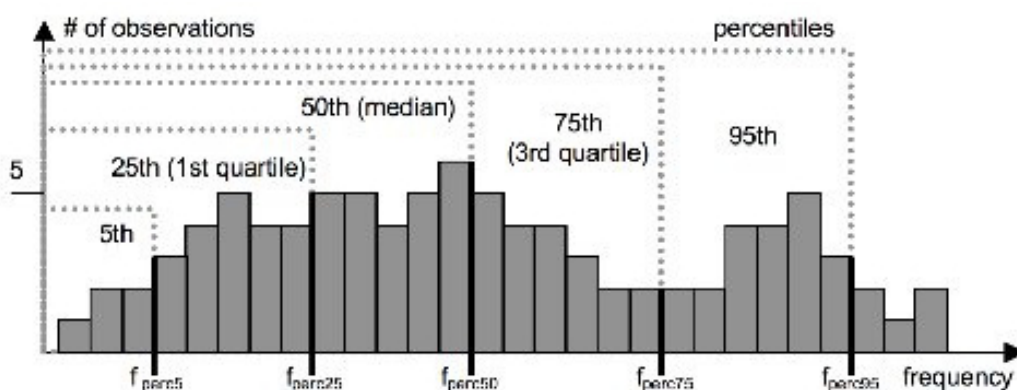
```
frame [6]:  
    intensity = 0.37551881872542725  
    nCandidates = 7  
    candidate []:  
        candidate [1]:  
            frequency = 171.7597654827709  
            strength = 0.8874489255215248  
        candidate [2]:  
            frequency = 489.4677208037172  
            strength = 0.3285399395373125
```

Az F0-értékek meghatározásának egyik fontos, a mérési eredményeket jelentősen befolyásoló paramétere, hogy az alapfrekvencia-értékek számításához milyen alsó („Pitch floor”) és felső („Pitch Ceiling”) küszöbértékeket választunk a számítás alapját képező tartomány meghatározásához a frekvenciaspektrum egészen belül. Az alsó küszöb a mérésekhez használt elemzési ablak időbeli hosszát határozza meg. Az érték növelésével rövidebb elemzési ablakot kapunk, így a Hertzben megadott értéknél kisebb frekvenciák kiesnek az elemzésből, a felső értéknél („Pitch Ceiling”) nagyobb frekvenciaértékeket pedig egyszerűen ignorálja az eredmény utófeldolgozása során. A Praat program dokumentációja²³ egy, a beszélők neme szerint meghatározott statikus beállítást (nők esetében 100 Hz-től 500 Hz-ig, férfiak esetében 75-től 300 Hz-ig) javasol. A két tartomány közötti automatikus választáshoz előzetes mérések alapján detektálnunk vagy egyéb formában rögzítenünk kell (pl. a fájlok nevében) a beszélők nemét. Az előbbi megoldás a HuComTech-korpusz esetében egészen megbízhatóan működött, ha az előzetes mérési tartományt valamivel szélesebbre (60 Hz-től 500 Hz-ig) állítva az egész fájlra számolt 150 Hz-es átlagos alapfrekvenciát határoztuk meg küszöbértéknek: a 150 Hertznel alacsonyabb átlagos alapfrekvenciájú felvételek az esetek többségében férfi, az annál magasabbak pedig női beszélők hangját rögzítették. Mivel a nemek megkülönböztetésére a hangmagasság alap-

²³http://www.fon.hum.uva.nl/praat/manual/Intro_4_2__Configuring_the_pitch_contour.html

ján nem alkalmazható biztonsággal efféle diszkrét határvonal²⁴, a sikeres (előzetesen rögzített) meghatározás esetén pedig jelentős egyéni eltérésekkel számolhatunk a nemeken belül, fontosnak találtam, hogy a fix beállítások helyett egy a beszélő egyedi hangterjedelméhez igazodó, dinamikus megoldást találjak a méréshez szükséges frekvenciatartomány meghatározásához, a statikus beállításokat opcióként hagyva meg.

Még egy korábbi tanulmány (Szekrényes, 2011) okán kísérletezni kezdtem egy, az F0-adatok eloszlásán alapuló módszerrel, amelyet Johan Frid és Gilbert Ambrazaitis a nemek szerinti standard beállítások elégtelen voltából kiindulva mutatnak be kapcsolódó tanulmányukban (Frid et al., 2010). A módszer lényege, hogy széles spektrumon (60 Hz-től 750 Hz-ig) végzett előzetes mérések alapján, statisztikai alapon próbál kiszámítani egy relatív, az adott beszélőre jellemző mérési tartományt, amelyben aztán az alaphfrekvencia végleges kontúrja meghatározható. Maguk a számítások az előzetes mérésekből származó, egy rendezett eloszlási mintában csoportosított F0-értékek kvantilis értékein nyugszanak. Az eloszlási minta kitüntetett részzeit egy másik, a fentebb említett szerzők munkáját megalapozó tanulmányból (Edlund & Heldner, 2007) kölcsönzött 4.12 ábra szemlélteti.



4.12. ábra. A frekvenciaértékek eloszlási mintája (Edlund & Heldner, 2007).

Frid és Ambrazaitis a módszerrel kapcsolatban De Looze & Hirst (2010) megelőző munkájára hivatkoznak, akik hasonló utat követve a frekvenciaeloszlás vizsgálatát a Praat lekérdező funkcióival oldották meg. Esetükben a végleges mérési tartomány alsó és felső határát az alábbi két, a Praat program beépített szkript nyelvén megfogalmazott képlet határozza meg:

$$\max F_0 = 10 * \text{ceiling}((1.92 * q_{65})/10) \quad (4.1)$$

²⁴ A beszélők biológiai neme a hangmagasság tekintetében sokkal inkább ragadható meg egy „elmosódott”, csak a fuzzy-logika eszközeivel megközelíthető kontinuum, mint éles terjedelmi határokkal bíró halmazok segítségével.

$$\min F_0 = 10 * \text{floor}((0.83 * q_{15})/10) \quad (4.2)$$

A felső határ számításához használt 4.1. képletben a q_{65} változó az előzetes mérési tartomány eredményeként kapott F_0 -adatok rendezett mintájából a 65. percentilis értéket tartalmazza, amely a 4.12. ábrán a 3. kvartilis érték és a medián között helyezkedik el. Az első kvartilis alatti 15. percentilis érték (a 4.2. képletben a q_{15} változó) pedig az alsó határ számításához szolgált. A korábban említett kísérletek (Szekrényes, 2011) során azt tapasztaltam, hogy a képletekben szereplő paraméterek bizonyos korrigációival javítható a mérések eredményessége. Az „eredményesség” alatt itt, mivel vizsgálatunk az alaphfrekvencia relatív ingadozására koncentráldódik, nem a pontosabb izolált értékek kinyerését vagy a meglévők elferdítését, hanem a relatív különbségek feltérképezéséhez szükséges adatok összegyűjtését értem, amit sok esetben megnehezített, ha a mérések hiányos vagy torz eredményekkel tértek vissza jellemzően az alsóbb frekvenciatartományok ignorálása, illetve a felső határ túlságosan magasra állítása esetén. A korrigálás érdekében már az előzetes mérési tartomány felső határát 250 Hertz-cel csökkentettem, hogy kizárjak néhány, a Praat néhol hibás F_0 -kalkulációiból (oktávugrások stb.) származó, extrém nagyságú frekvenciaértéket. A végleges felső határ számításához pedig visszatértem a de Looze és Hirst által korábban szintén használt 3. kvartilis értékhez az alsó határ keresésében egészen az 5. percentilis értékig jutva. A végeredményül kapott művelet sor a Praat beépített szkript nyelvére átfordítva:

```
To Pitch... 0.01 60 500
q05 = Get quantile... 0.0 0.0 0.05 Hertz
q75 = Get quantile... 0.0 0.0 0.75 Hertz
min_f0 = 10*floor((0.93*q05)/10)
max_f0 = 10*ceiling((1.5*q75)/10)
```

Az eredményül kapott $\min_f0$ és $\max_f0$ változók az F_0 -adatok végleges kinyerése során így dinamikusan, a beszélő egyedi (az előzetes mérésből származó) hangmagasság értékeinek eloszlásához igazodva határozzák meg a számításhoz használt frekvenciatartomány alsó és felső határát. A mérések természetesen ismét a beszélő izolált hangját tartalmazó hangfelvételen hajtnak végre a Praatban alapértelmezettként beállított 10 milliszekundumos léptetéssel:

```
To Pitch... 0.01 min_f0 max_f0
```

A ProsoTool a fent bemutatott módszerekkel mért alaphfrekvencia-adatok segítségével végzi el a beszélő egyedi hangterjedelmének vizsgálatát. A mérés eredményeként létrejött „Pitch”

objektumban található F0-értékek átlaga („Get mean”) és szórása („Get standard deviation”) alapján négy küszöbérték kerül definiálásra:

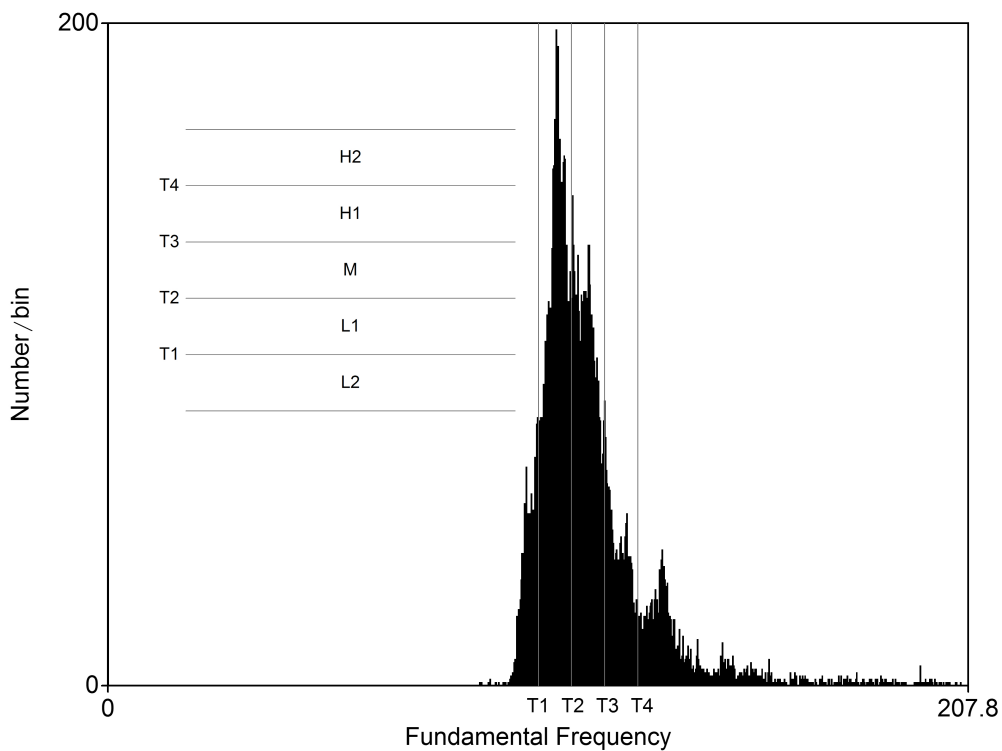
$$T_1 = \langle F_0 \rangle - \sigma(F_0) \quad (4.3)$$

$$T_2 = \langle F_0 \rangle - \frac{\sigma(F_0)}{3} \quad (4.4)$$

$$T_3 = \langle F_0 \rangle + \frac{\sigma(F_0)}{3} \quad (4.5)$$

$$T_4 = \langle F_0 \rangle + \sigma(F_0) \quad (4.6)$$

A küszöbértékek a beszélő egyedi hangterjedelmét öt, a hangmagasság különböző tartományainak megfeleltethető részre tagolják.

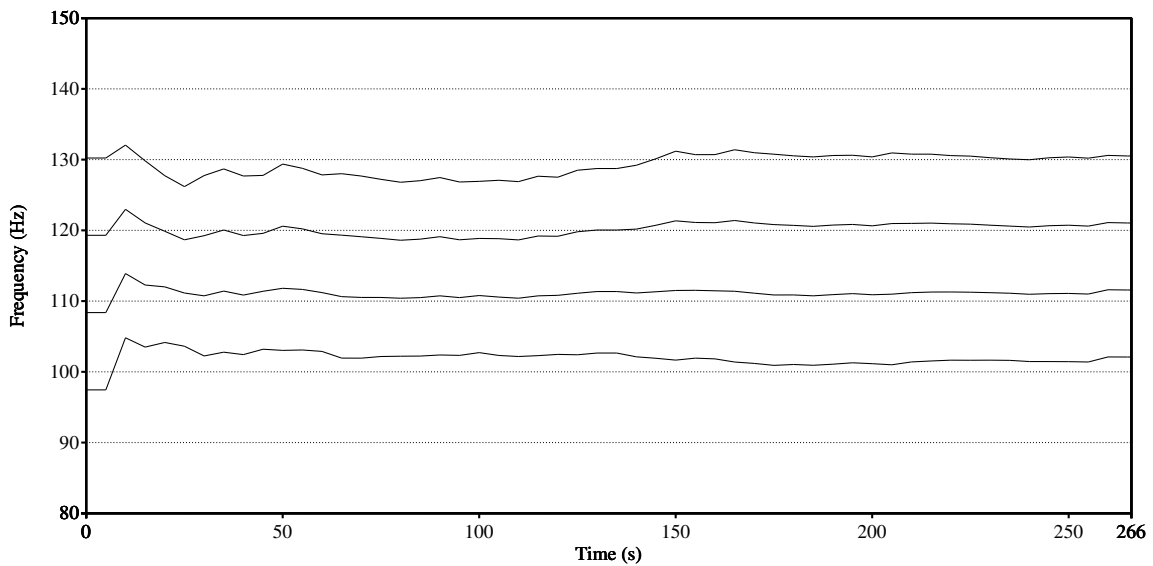


4.13. ábra. A beszélő vokális hangterjedelmének felosztása az F0-adatok eloszlása alapján.

A 4.13. ábra az alaphékvencia-értékek eloszlását, ugyanezen értékek átlaga és szórása alapján számolt küszöbértékeket, illetve a küszöbértékek által meghatározott (a 4.2.1.2. szakaszban már bemutatott) tartományokat szemlélteti, melyek közül az L_2 az alaphang legalacsonyabb, a H_2 pedig a legmagasabb szintjeit jelöli:

$$\mathbf{L}_2 < T_1 < \mathbf{L}_1 < T_2 < \mathbf{M} < T_3 < \mathbf{H}_1 < T_4 < \mathbf{H}_2 \quad (4.7)$$

A küszöbértékek számításakor természetesen nem mindegy, hogy milyen időbeli hosszúságú hangfelvételt veszünk alapul. A 4.14. ábrán látható, hogy a példaként kiválasztott beszélő esetén ≈ 50 másodpercen belül az alaphfrekvencia-értékek átlaga és szórása még elég sokat ingadozhat, azonban egy idő után ($t > 150s$) ez az ingadozás radikálisan mérséklődni kezd, ezáltal a küszöbértékek pozíciója is egy bizonyos frekvenciaérték mentén állandósul.



4.14. ábra. A küszöbértékek változása a számításhoz alapul vett hanganyag időbeli hosszának függvényében. A görbék pontjai azt fejezik ki, hogy az adott időpillanatig mért alaphfrekvencia-értékek alapján a különböző küszöbértékek hova esnének a beszélő hangterjedelmében.

4.2.1.5. Az F0-adatok feldolgozása

A beszélő egyedi hangterjedelmének feltérképezése után az algoritmus további műveleteinek bemenete már nem beszélő izolált hangját tartalmazó vágott anyag, hanem az eredeti hangfelvétel. Az F0-adatok elemzése azonban szintén a választott beszélő megnyilatkozásainak az előfeldolgozás során megjelölt szünettől szünetig, szünettől beszélőváltásig (vagy fordítva), illetve beszélőváltástól beszélőváltásig tartó intervallumain belül történik: szegmentumról szegmentumra haladva.

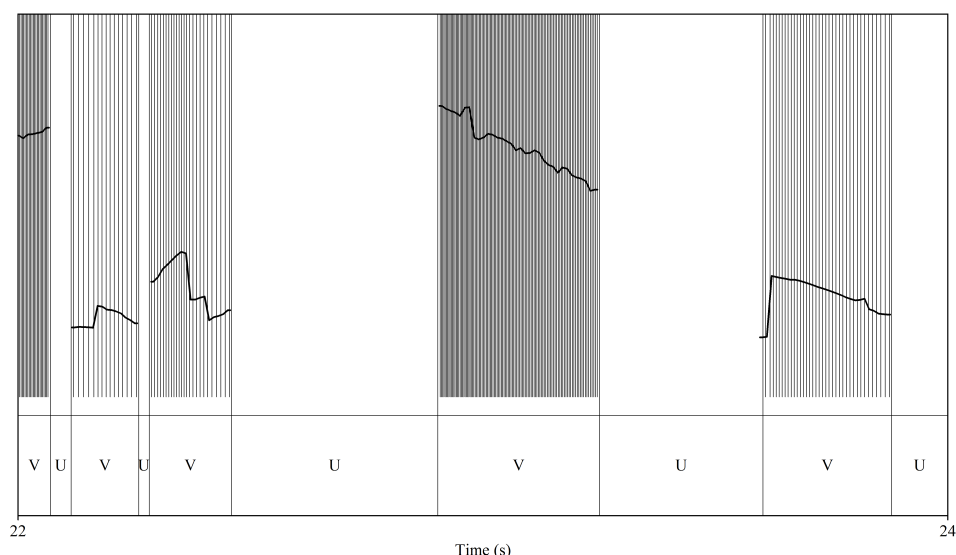
A szegmentumonként végzett elemzés első lépése ismét a nyers F0-értékek kinyerése lesz, amely az előző szakaszban, a beszélő teljes hanganyaga alapján megállapított, a `min_f0` és a `max_f0` változóban tárolt értékek által határolt frekvenciatartomány használatával zajlik. A Praat program parancsai közül itt most nem a „To Pitch”, hanem több paraméterrel hangolható „To Pitch (ac)” utasítást építettem bele az algoritmusba az alábbi beállítások mellett:

To Pitch (ac): 0.01, min_f0, 15, "no", 0.045, 0.45, 0.01, 0.35, 0.14, max_f0

A beállított értékek sorrendben az alábbi, Praatban használatos paramétereknek²⁵ feleltek meg: „Time step” (a mérésekhez használt léptetést, a mérési helyek sűrűségét határozza meg), „Pitch floor” (az F0 számításához figyelembe vett legkisebb frekvencia, az érték az elemzési ablak méretére hat), „Maximum number of candidates” (a lehetséges alapfrekvencia-értékek maximális száma), „Very accurate” (logikai kapcsoló, mely az elemzési ablak két lehetséges típusa között vált), „Silent threshold” (a beszédsszünetek detektálásához használt küszöbérték), „Voicing threshold” (a zöngés szegmentumok detektálásához használt küszöbérték), „Octave cost” (a magasabb frekvenciaértékek előnyben részesítésének foka), „Octave jump cost” (az oktáv ugrások kiküszöbölésére szolgál), „Voiced/unvoiced cost” (a zöngés/zöngétlen közötti választást befolyásolja: az érték növelésével a zöngésnek ítélt szegmentumok száma csökken), „Pitch ceiling” (a beállított értéknél nagyobb frekvenciákat ignorálja). Az alapértelmezett beállításokhoz képest a „Silent threshold” értékét a beszédsszünetek javára növeltem meg, hogy a magasabb frekvenciájú, de gyengébb intenzitású zörejekhez (főként bizonyos mássalhangzók esetén fordultak elő) a Praat ne rendeljen tévesen az elemzés torzulását kiváltó (kiugróan magas) alapfrekvencia-értékeket.

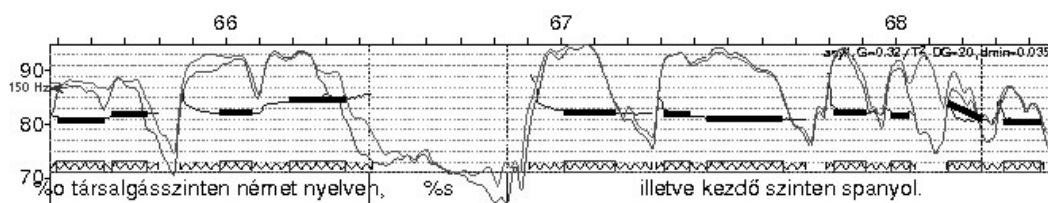
Az algoritmus következő lépése a zöngés és zöngétlen beszédsszakaszok detektálása és annotációja az adott szegmentumon belül. A műveletek végrehajtására a Praat program beépített utasításait használtam fel. A „Pitch” típusú objektumok esetén elérhető „To PointProcess” parancs („Analyze” utasításcsoport) egy, a hanghullám kváziperiodikus elemeinek helyét jelölő „PointProcess” típusú objektumot generál, ahol a periódusok sűrűsödése és ritkulása az alapfrekvencia emelkedését és csökkenését reprezentálja. A „PointProcess” objektum adatai alapján a „To TextGrid (vuv)” parancs egy „TextGrid” objektumban a beszédet automatikusan zöngés és zöngétlen szakaszokra szegmentálja, amelyek aránya később a dallamformák osztályzásánál nyer igazi jelentőséget, de mint részeredmény a prozódiai annotáció kimenetének is részét képezi. A Praat vonatkozó objektumainak („Pitch”, „PointProcess”) viszonyait és az eredményül kapott annotációt a 4.15. ábra szemlélteti, ahol a vertikális vonalak a periódusok sűrűségét és az alapfrekvencia (fekete görbe) ezzel egyenes arányos magasságát ábrázolják. Az annotációban az „U” („Unvoiced”) szimbólum jelöli a zöngétlen, a „V” („Voiced”) pedig a zöngés beszédrészeket.

²⁵ http://www.fon.hum.uva.nl/praat/manual/Sound__To_Pitch__ac____.html



4.15. ábra. Zöngés és zöngétlen szakaszok detektálása a Praat programmal.

A ProsoTool korábbi verzióiban (Szekrényes, 2014) a nyers F0-értékek stilizálást a 3. fejezetben már bemutatott Prosogram alkalmazás beépítésével: külső Praat szkriptként való meghívásával és szegmentumonként történő lefuttatásával végeztem el.²⁶ A dallamformák osztályozása ekkor a Prosogram „PitchTier” formátumban tárolt köztes eredményei (az alaphérfekvencia szótagszinten, értékpárok formájában stilizált vektorai) alapján történt. A stilizálás grafikus kimenetét a 4.16 ábra szemlélteti.



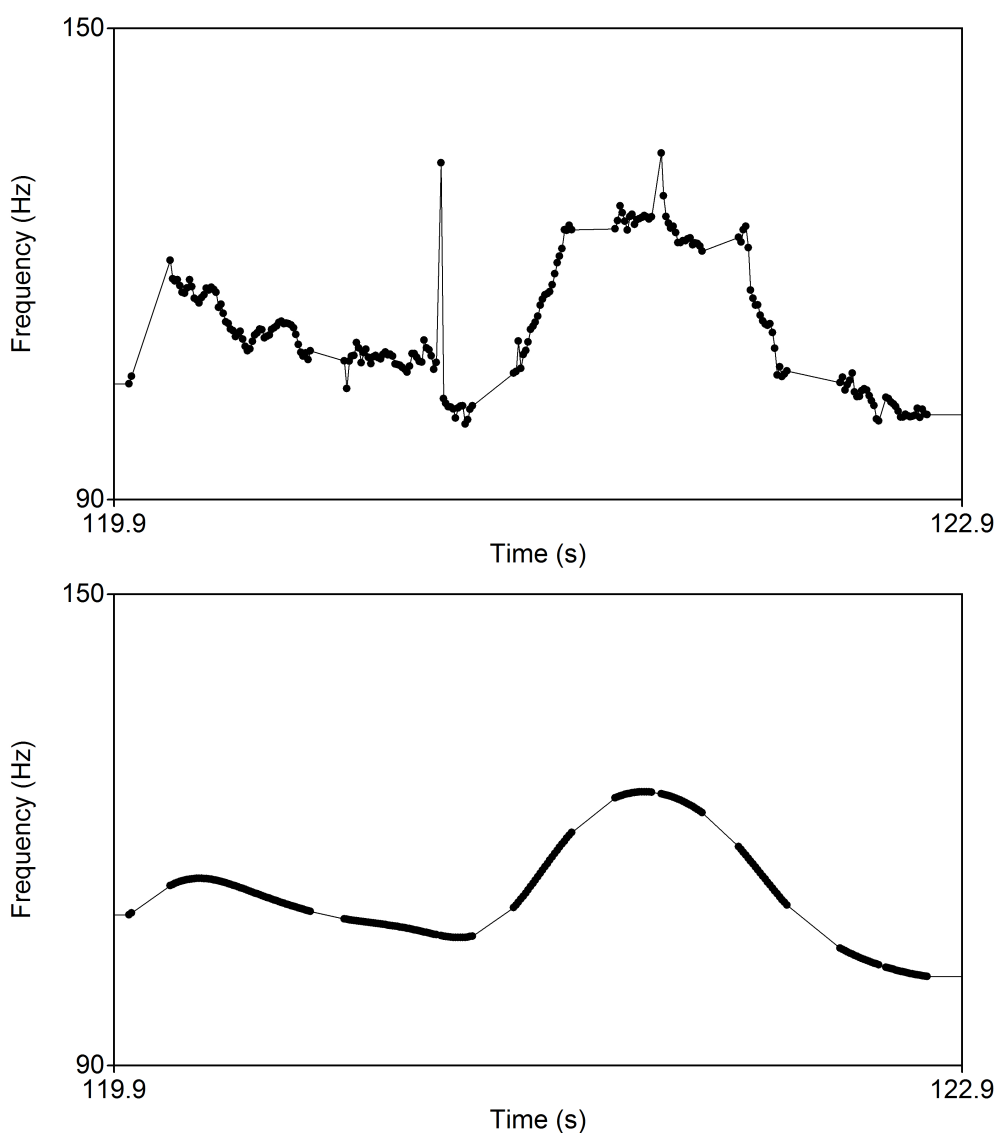
4.16. ábra. Az F0-görbe stilizálása a Prosogram segítségével (korábbi verziókban).

A beszéddallam ezen módszerekkel nyert, szótagszintű reprezentációjáról azonban hamar kiderült, hogy habár fonetikai értelemben egy rendkívül részletgazdag anyag birtokába jutunk, a Prosogram főntebb illusztrált kimenete a kommunikációs funkciók feltárásának és elemzésének szempontjából nem felel meg a HuComTech korpusszal kapcsolatos további kutatások igényeinek. A ProsoTool korábbi verzióiban a szótagszintű stilizációkat hosszabb, akár több szótagot is egyetlen dallameseményben integráló alaphérfekvencia-trendekben integráltam. Problémát jelentett viszont, hogy a Prosogram eredményeinek formátuma („PitchTier” típusú Praat objektum) a Praat programban nem támogatta sem az F0-értékek simítását, sem a zöngés-zöngétlen

²⁶ Ezekben a verziókban az alaphérfekvencia szegmentumonkénti mérését is a Prosogramnak engedtem át, az előző szakaszban tárgyalt paraméterek átadásával.

szakaszok főtebb bemutatott detektálását. A „Pitch” objektumokból leképezhető „PitchTier” objektum ugyanis egy olyan egyszerűsített reprezentációs forma, ami nem őrzi meg az ehhez szükséges adatokat, csak a legvalószínűbb frekvenciaértékeket és azok időzítését tartalmazza.

A nyers F0-értékeken a Prosogram stilizáló eljárása helyett a ProsoTool jelenlegi verziójában (Szekrényes, 2015) a feldolgozás következő lépéseként a Praat program beépített, az alaphfrekvencia görbe simítására szolgáló algoritmus (a „Convert” utasításcsoport „Smooth” parancsa) kerül lefuttatásra, amelynek eredményét a 4.17. ábra szemlélteti.



4.17. ábra. A F0-görbe simítása („Bandwidth” paraméter: 1,5 Hz). Eredeti görbe: fent. Simított görbe: lent.

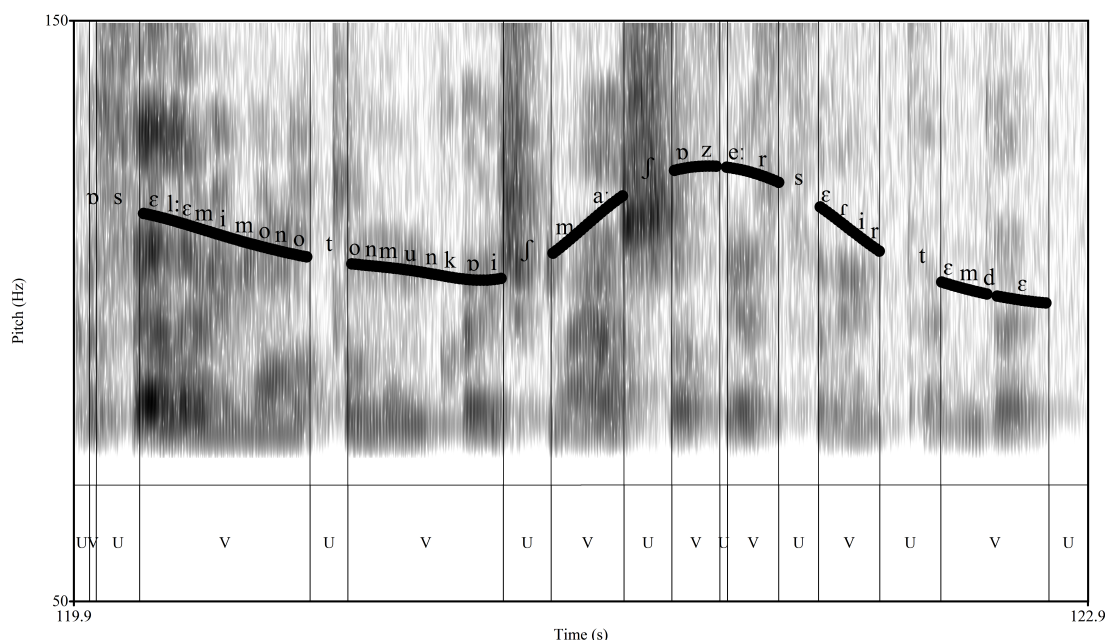
Magáról az algoritmusról a Praat program egyéb tekintetben rendkívül alapos és terjedelmes dokumentációjában meglehetősen szűkszavúan fogalmaznak.²⁷ Csak egy kapcsolódó internetes

²⁷ http://www.fon.hum.uva.nl/praat/manual/Pitch_Smooth_.html

levelezőlistán (*Yahoo! Groups: Praat Users Group*) találhatunk rá a szoftver rendszeres fejlesztését végző Paul Boersma valamivel részletesebb leírására:

„...in the time domain, it's a convolution with a Gaussian with the said bandwidth. That is, a multiplication in the frequency domain (of the Fourier-transformed pitch curve) with a shape of $\exp(-(frequency/bandwidth)^2)$. Before the filtering, the voiceless stretches are filled up with voicing by linear interpolation and (at the edges) constant extrapolation; after the filtering, these parts are made voiceless again...”²⁸ (Boersma 2011)

Az alapfrekvencia-görbe simítása kettős haszonnal jár: sok esetben kiküszöböli a Praat alapfrekvencia-számításának esetleges hibáit (például a 4.17. ábra felső részén, a 121. másodperc környékén látható kiugróan magas F0-értéket), továbbá a simítás eredményeként nyert görbe az F0-értékek a percepció számára irreleváns, néhány Hertz értékű bukdácsolását már nem fogja tartalmazni. A művelet nyilvánvaló információvesztéssel is jár a görbe eredeti formájára vonatkozóan, a stilizálás szempontjából viszont kifejezetten előnyös, ha az értékek kevésbé fontos változásait figyelmen kívül hagyjuk, és a 4.17. ábra alsó részében látható markánsabb és laposabb dallamívekben integráljuk.



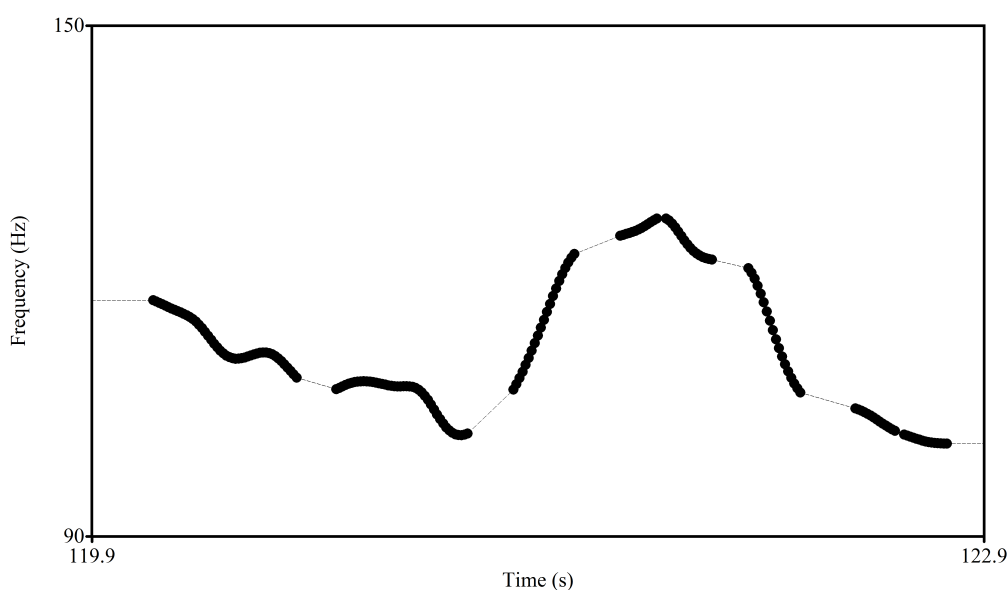
4.18. ábra. A simított F0-görbe és hozzá tartozó beszédszakasz kváziperiodikus szegmentumainak pozíciója. A háttérben a beszéd részlet hangszínképe látható 0–5000 Hertzes tartományban. A felvételen az alábbi megnyilatkozás-részlet hallható (a simított görbe fölött ugyanez fonetikus átírásban szerepel): „...a szellemi monoton munka is más azér’ szerintem, de...”.

²⁸ <https://uk.groups.yahoo.com/neo/groups/praat-users/conversations/topics/4168>

A 4.18. ábrán a simított F0-görbe és a hozzá tartozó beszédszakasz kváziperiodikus (az ábrán „V”-vel jelölt zöngés intervallumok) szegmentumainak pozíciója látható, amelyeket – mint Paul Boersma fentebbi leírásában is olvasható volt – a simítás eredményeként létrejövő „Pitch” objektum továbbra is megőriz. Az ábrán vizualizált beszédszakasz²⁹ az alábbi megnyilatkozást tartalmazza:

A beszédrészlet fekete vonalakkal jelölt dallamának folytonossága a zöngétlen mássalhangzók és a beszédszünetek helyén megszakad. A 4.18. ábrán ezt az [s], [t], [ʃ] zöngétlen mássalhangzóknál figyelhetjük meg.³⁰ A zöngés szegmentumokon belül található dallamívek pedig akár több magánhangzót és szótagot is tonális egységekbe fognak.

Mint a ProsoTool algoritmusának rövid leírásában (4.2.1.2. fejezet) olvashattuk, lehetőségünk van a simítás mértékét meghatározó „Bandwidth” paraméter Hertzben megadható értékének megváltoztatására. A paraméter értékének növelése enyhébb, csökkentése erősebb simítást eredményez. A 4.19. ábrán az előző példában szereplő beszédrészlet alapfrekvencia-értékeinek feldolgozásánál az enyhébb simítás érdekében a paraméter értékét 5 Hertzre növeltük. A ProsoTool forráskódjában 1,5 Hertz az alapértelmezés.



4.19. ábra. Az F0-görbe simítása. „Bandwidth” paraméter értéke: 5 Hertz.

²⁹ Forrás: HuComTech korpusz, 003mv19_F azonosítójú formális interjúfelvétel, 119.9 – 122.9 mp.

³⁰ A „munka” szóban, az [a] és az [n] között szereplő [k] hang a Praat program számításai alapján kváziperiodikus elemnek minősült.

4.2.1.6. A dallamformák stilizálása, szegmentálása és osztályozása

Az algoritmus következő lépése a simított alapfrekvencia-görbe stilizálása és ún. *intonációs eseményekre*³¹ való szegmentálása. A stilizálás végrehajtásához a Praat program „Stylize” funkcióját használtam fel, ami a „Pitch” objektumban tárolt adatsorok az egyszerűsített „PitchTier” formátumra való konvertálása („Down to PitchTier” parancs a „Convert” csoporton belül) után a dinamikus menü „Modify” utasításcsoportjában érhető el. A stilizálás a „Frequency resolution” paraméterrel szabályozható. A paraméter értékét Hertzben kifejezve, vagy félhangértékekben relatív módon is megadhatjuk. Az előző szakaszban bemutatott alapfrekvencia-görbén elvégzett stilizálás eredményét szemlélteti a 4.20. és a 4.21. ábra eltérő frekvencia felbontások használatával. A művelet lépéseit a Praat program online dokumentációja³² az alábbi pontokban foglalja össze:

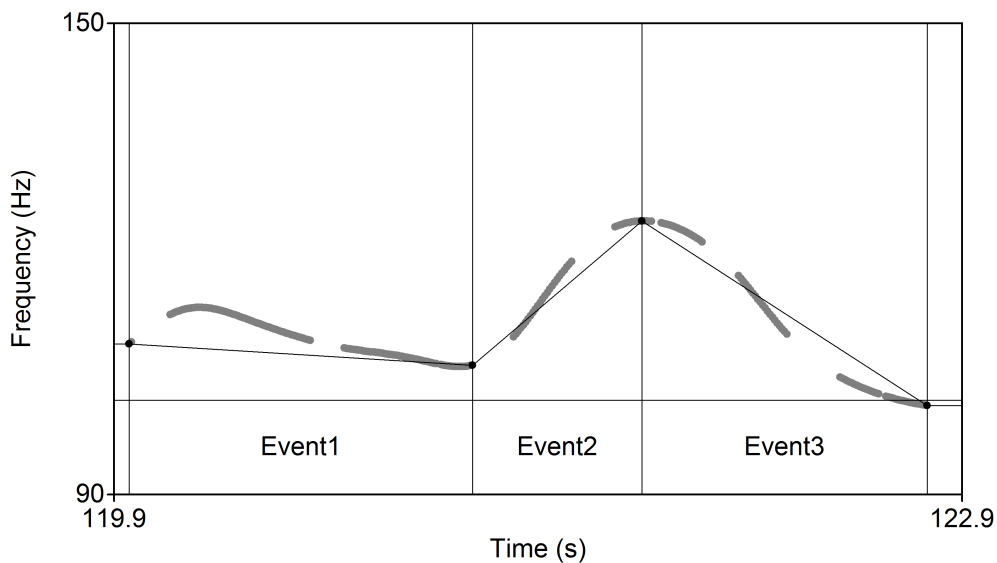
1. Keresd meg azt a sorrendben következő alapfrekvencia-értéket, ami a legközelebb esik, ahhoz az „egyeneshöz”, ami a vele szomszédos értékeket összeköti egymással.
2. Ha ez az érték távolabb helyezkedik el az egyenestől mint a frekvencia-felbontás megadott értéke („Frequency resolution” paraméter), akkor az alapfrekvencia-görbe már nem stilizálható tovább
3. Ha a távolság kisebb (vagy a felbontással egyező), akkor az első lépésben talált értéket eltávolítjuk.
4. Térj vissza az első lépéshez.

A stilizálás eredményeként a simítás során átdolgozott F0-görbék még tovább egyszerűsödnek a beállított frekvencia-felbontás függvényében.³³ A simítással ellentétben a Praat stilizáló eljárása nem pusztán a görbe formájára van kihatással, de mint az algoritmus leírásában is olvasható: bizonyos, az alapfrekvencia kisebb változásait jelző értékek eldobásra is kerülnek, csak a jelentősebb (a beállított felbontás értékét meghaladó) eltéréseket reprezentáló értékek maradnak meg. A ProsoTool algoritmusában a stilizálás funkciója a simítás eredményeként kapott dallamformák az alaphang lényegesebb változásait követő intonációs eseményekre való szegmentálása, amelyeket a 4.20. és a 4.21. ábrán a stilizált görbe frekvenciahelyeit összekötő („*Event_n*” címkével jelölt) trendvonalak vizualizálnak.

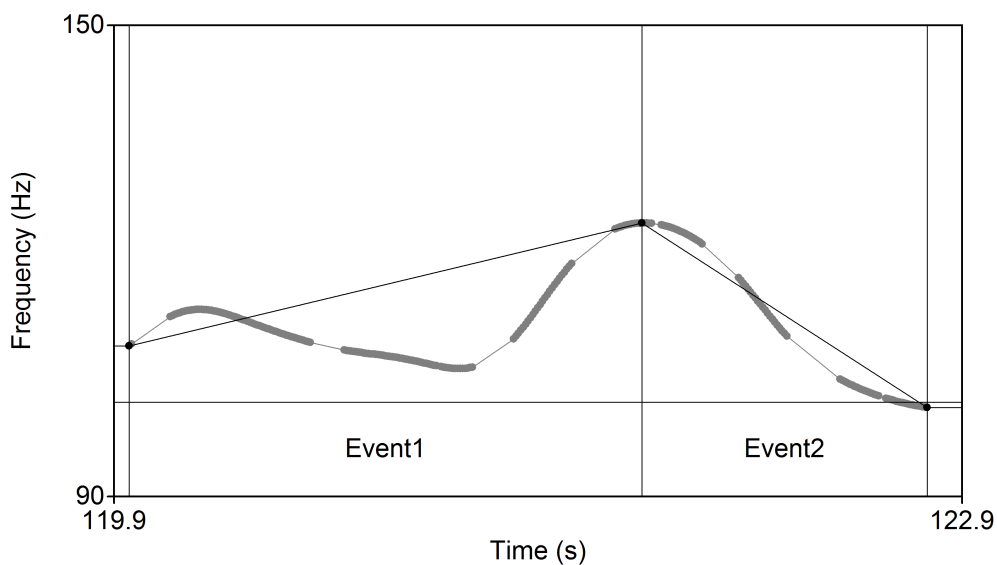
³¹ Az elnevezés a Tilt intonációs modell (Taylor, 2000) terminológiáját („intonational events”) követi.

³² http://www.fon.hum.uva.nl/praat/manual/PitchTier__Stylize___.html

³³ Ez a paraméter értékének növelésével egyenes arányosságban történik. A simítás „Bandwith” paramétere esetén az arány fordított volt.



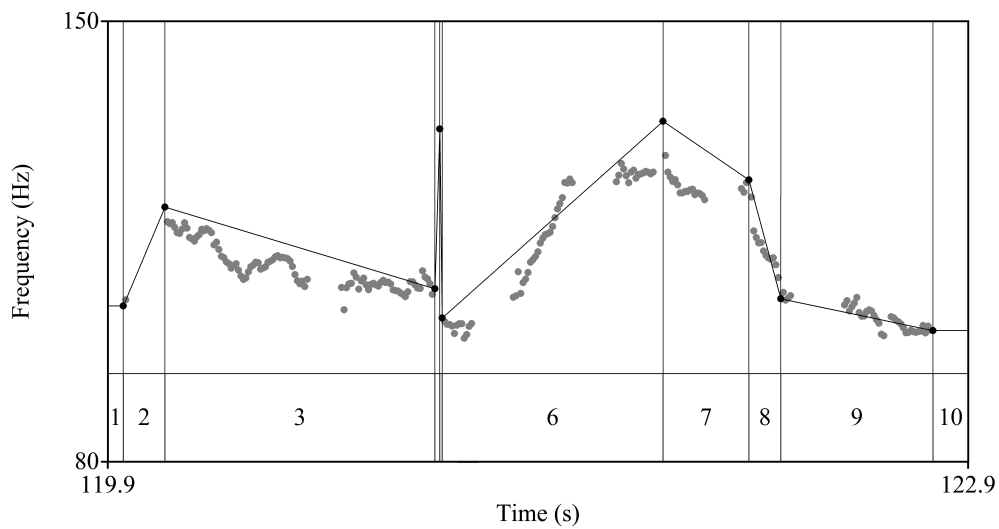
4.20. ábra. Az F0-görbe stilizálása (frekvenciafelbontás: 10 Hz).



4.21. ábra. Az F0-görbe stilizálása (frekvenciafelbontás: 20 Hz).

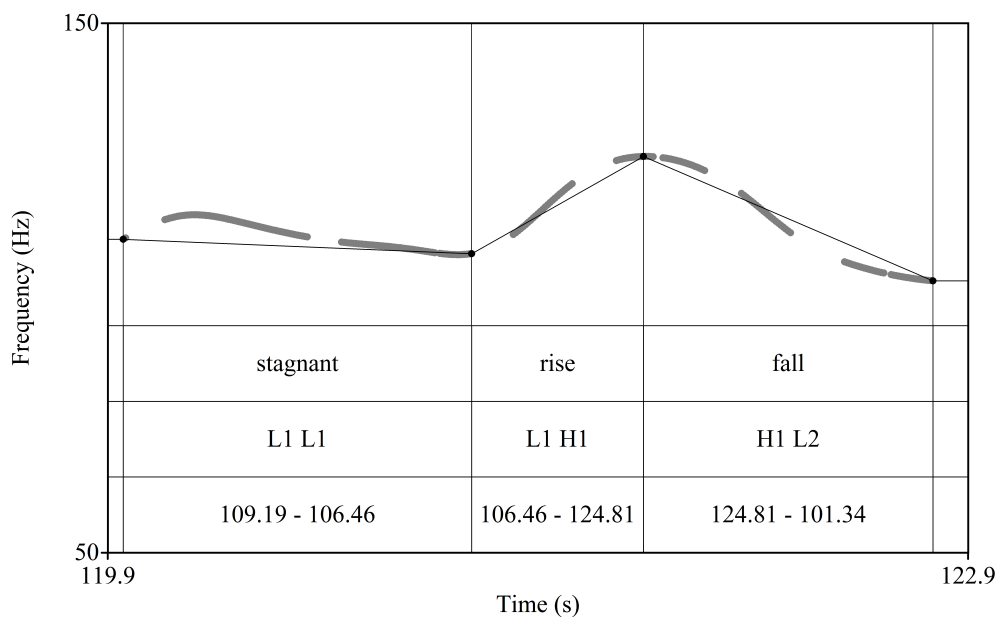
A 4.20. ábrán „Event1”, „Event2” és „Event3” címkékkel jelölt, a stilizálás eredményeként kapott trendvonalak az alaphfrekvencia változásait három fontosabb intonációs eseményre szegmentálják, ahol a szegmentumok akár több, zöngétlen szakaszokkal tagolt kváziperiodikus intervallumot is ugyanazon tendencia részeként egyesítenek. Maga a szegmentálás a stilizálás után megmaradt F0-értékek mentén történik. A frekvenciafelbontás további növelésével (lásd 4.21. ábra) a szegmentumok száma csökkenni fog. Ha a stilizálást a példaként használt alaphfrekvencia előzetes simítása nélkül végeztük volna el, akkor az intonációs események a 4.22. ábrán látható, a beszéddallam elemzése szempontjából előnytelen, rövid időtartamú és nagy-

számú szegmentumaihoz jutunk (a 10 Hertz-es frekvenciafelbontás megtartása mellett), amelyet a percepció ítéletei is alacsony valószínűséggel igazolnának vissza.



4.22. ábra. Az F0-görbe stilizálása előzetes simítás nélkül (frekvenciafelbontás: 10 Hz).

A ProsoTool algoritmusának utolsó lépését a stilizálás alapján létrehozott intonációs események szabályalapú osztályozása jelenti. A 4.23. ábrán, a stilizált alapfrekvencia-görbe alatt látható, az elemzés eredményeként generált, az események szegmentumait követő címkesorok a 4.2.1.2. fejezetben már ismertetett annotációs szintekből állnak.



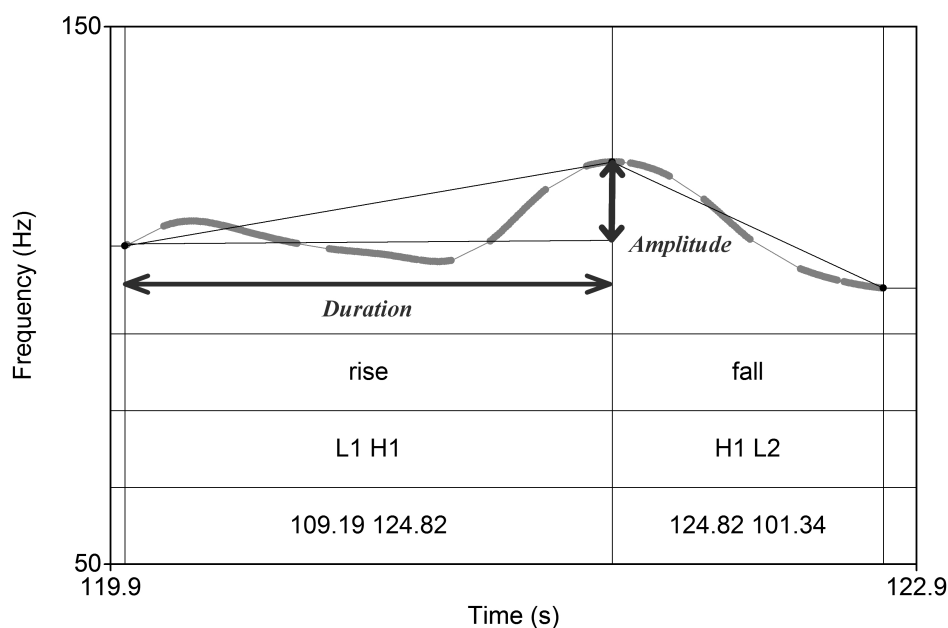
4.23. ábra. Az intonációs események jellemzése és osztályozása.

A legelső annotációs szinten az intonációs események elején és végén mérhető alapfrekvencia-értékek szerepelnek Hertzben megadva. A második szinten ugyanezek az értékek a 4.2.1.4. szá-

kaszban bemutatott módszerrel felosztott egyéni hangterjedelem megfelelő szintjein ($L_2...H_2$) kerülnek elhelyezésre. Az alaphang így kifejezett, relatív „mozgása” a beszélő egyéni hangterjedelmén belül már önmagában informatív lehet a beszéddallam időbeli modulációira, az intonáció strukturális szerkezetére vonatkozóan. A prozódia későbbi elemzése szempontjából azonban hasznosnak ígérkezett, ha az egymást követő intonációs események kontúrjáról az annotáció explicit információkat is tartalmaz, mint a 4.23. ábra legfelső szintjén látható. A dallamformák osztályozása nem az alapfrekvencia második szinten jelölt relatív viszonyain, hanem részben a Tilt intonációs modellben (Taylor, 2000) is használt RFC paramétereken (lásd 3.3.2. fejezet), részben pedig a beszélő izolált hangja alapján mért alapfrekvencia-értékek eloszlásán alapul. Az osztályozást végző eljárás kimenetét három fontos, a modulációra vonatkozó (a 4.24. ábrán is szemléltetett) paraméter határozza meg:

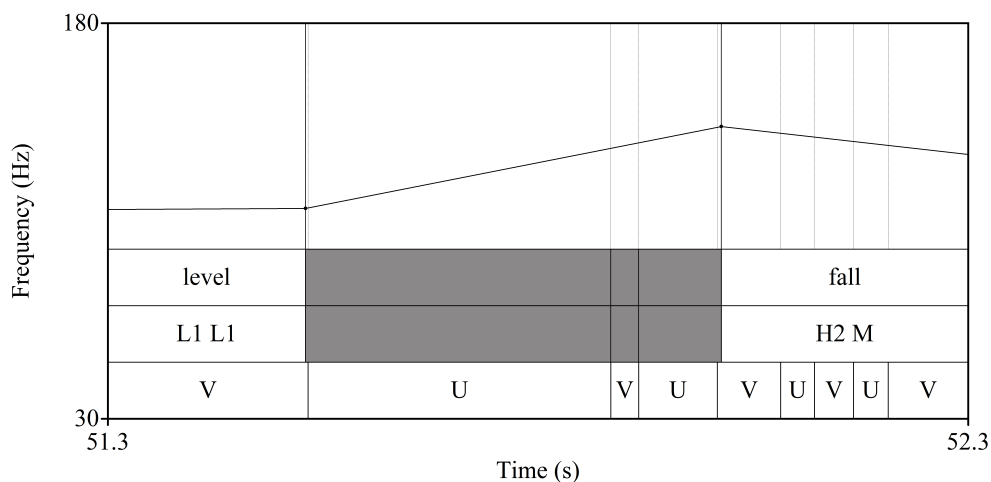
1. a moduláció amplitúdója,
2. a moduláció időtartama,
3. a moduláció iránya (előjele).

A fenti paraméterek kerülnek összehasonlításra a beszélő egyéni hangterjedelmén belül egy szintnek megfelelő, az alapfrekvencia-értékek szórása alapján meghatározott, az osztályzó eljárásban küszöbértékként használt frekvenciatartománnyal. Ha a moduláció amplitúdója ezt a küszöbértéket meghaladja, akkor az alaphang a beszélő hangterjedelmén belül legalább egy szintet ugri (például az M szintről a H_1 szintre ér), így a moduláció időtartama (gyorsasága) és iránya függvényében az osztályzás kimenete az alábbi címkék valamelyike lesz: „*rise*” (a magyar terminológiai megfelelő: *szökő*), „*ascending*” (*emelkedő*), „*descending*” (*ereszkedő*), „*fall*” (*eső*). Minden más esetben „*level*” (*szinttartólebegő*) címkét kapunk eredményül. Az intonációs esemény időtartama az „*ascending*” és a „*rise*”, illetve a „*descending*” és a „*fall*” címkék elkülönítésében számít fontos tényezőnek.



4.24. ábra. A dallamformák osztályzásánál figyelembe vett paraméterek.

Mivel a stilizált dallamformákat tartalmazó „PitchTier” objektum a köztes (a trendeken belüli) alapfrekvencia-értékekre vonatkozó információkat már nem tartalmaz, az intonációs események osztályzása előtt a zöngés és zöngétlen szegmentumok – az előző szakaszban bemutatott – annotációja fontos szerepet kap az intonációs szünetek helyét jelölő, valódi modulációt nem tartalmazó intervallumok kiszűrésében. A 4.25. ábrán szürke színnel jelölt intervallum (egy „level” és egy „fall” címkével jelölt intonációs esemény között) a zöngés („V”) és zöngétlen („U”) szegmentumok aránya ($V : U < 3 : 7$) alapján például egy intonációs szünet helyét jelöli, amely az eseménysor osztályzásából kimaradva megszakítja a beszéd dallammenetét, ami a következő esemény kezdetén a H_2 szinten folytatódik.

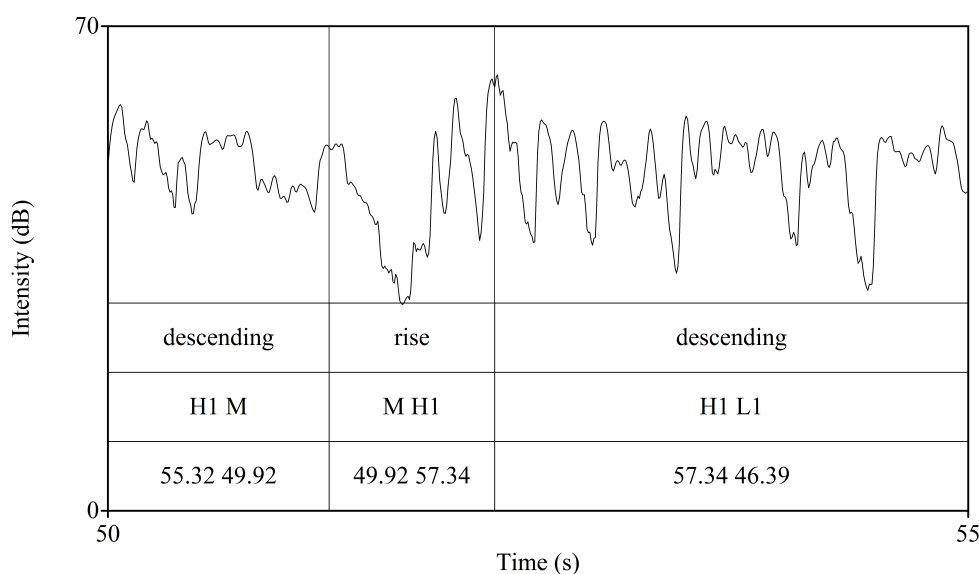


4.25. ábra. Intonációs szünetek detektálása.

4.2.2. Az intenzitás annotációja

A beszédintenzitás annotációjához a fontosabb lépések tekintetében a beszéddallam esetén használt algoritmus került ismételt alkalmazásra, amely bizonyos átalakítások után a decibel értékekben kifejezett intenzitásgörbék feldolgozására is használhatónak bizonyult. Az előfeldolgozás 4.2.1.3. alfejezetben bemutatott módszerei és bemeneti követelményei teljesen azonosak, a beszéddallam annotációjához előkészített anyagon a beszédintenzitás hasonló feldolgozása is elvégezhető.

Az annotáció kimenetének (lásd 4.26. ábra) formája és szerkezeti felépítése is megegyezik a beszéddallam esetén generált struktúrával: a jelentősebb modulációkat kifejező trendekben integrálja a beszéd intenzitásának változásait, amelyeket abszolút (decibelben) és relatív (a beszélő hangjának egyedi sajátosságai alapján megállapított intenzitás szintek) módon kifejezett értékpárok formájában, illetve az intenzitásgörbe kontúrjára vonatkozó címkékkel jellemez. A relatív intenzitás szintek megállapítása a beszélő izolált hangjából mért intenzitásértékek eloszlása alapján történik a 4.2.1.4. szakaszban az F0-adatokon már bemutatott módszerekkel.

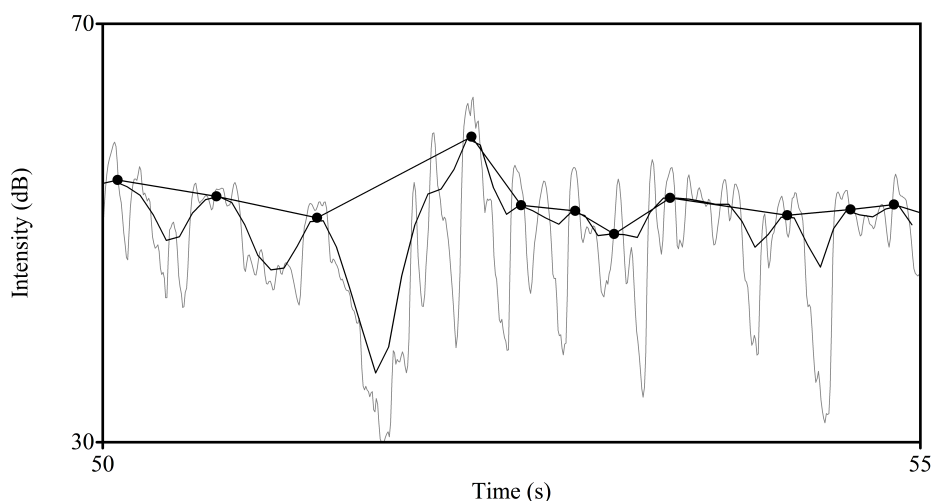


4.26. ábra. A beszédintenzitás annotációjának kimenete.

A beszédintenzitás szegmentumonkénti (a beszélőváltás korábban bemutatott akusztika szerkezetét követő) feldolgozásához eltérő technikai megoldásokat tett szükségessé az intenzitásértékek gyakori, a beszédhangokkal szinkronizált, de a percepció számára irreleváns ingadozása. Mivel az alaphétfrekvencia-görbékre alkalmazott „Smooth” függvény az „Intensity” típusú objektumok esetén a Praat programban nem elérhető, az intenzitásgörbe simítása már a nyers adatok

mérésekor a „To Intensity” parancs „minimum Pitch” paraméterének beállításával³⁴, majd a görbe csúcsértékeinek („To IntensityTier (peaks)”) kinyerésével volt megvalósítható.

A 4.27. ábrán látható halványszürke görbe a „To Intensity” parancs a 4.26. ábrán már bemutatott kimenetét szemlélteti ugyanazon az 5 másodperces beszédszakaszon. Futtatás előtt a „minimum pitch” paraméter értékét az alapértelmezett 100 Hertzre állítottuk. A beszédintenzitás szótagok közötti ingadozását kevésbé részletgazdagon kifejező, csak a markánsabb tendenciákat megragadó fekete görbe ugyanezen beszédrészlet intenzitásviszonyait ábrázolja a „minimum pitch” paraméter 10 Hertzre csökkentése mellett. A maximum helyeket összekötő burkológörbe a „To IntensityTier (peaks)” parancs konverziójának eredménye, amely az algoritmus következő lépésében a beszédintenzitás stilizálásának bemenete lesz.



4.27. ábra. Az intenzitás értékek feldolgozása a stilizálás előtt.

Mint a 4.27. ábrán látható, a csúcsértékeket összekötő görbe már stilizálás nélkül is figyelmen kívül hagyja az intenzitásnak a szótagmagok között megfigyelhető (a magánhangzók magasabb energiaszintjéből fakadó), de a beszéd percepciója szempontjából nem releváns csökkenését.

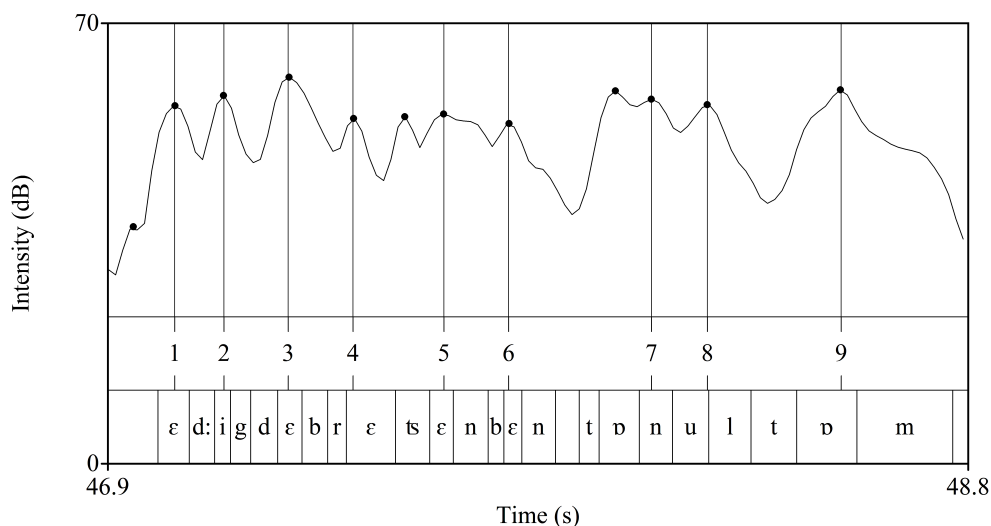
4.2.3. A beszédtempó annotációja

A beszédtempó a beszéd egy adott időegységre (perc, másodperc) eső szegmentális egységeinek (hang, szótag) *átlagos* mennyiségével ragadható meg, aminek kifejezésére hagyományosan két jellemzőt: a beszéd sebességét (szótagszám / a beszéd teljes időtartama) és az artikuláció sebességét (szótagszám / artikulációra fordított idő) különböztetik meg. A jellemzők egy

³⁴ http://www.fon.hum.uva.nl/praat/manual/Intro_6_2__Configuring_the_intensity_contour.html

másik aspektusból tovább csoportosíthatók globális, illetve lokális beszédtempóra. Az előbbi a teljes beszédanyagra számolt átlagokról ad számot, ami főként a különböző kommunikációs helyzetekben (hírolvasás, spontán beszéd stb.) megfigyelhető tempóbeli eltérésekre és azok percepciójára koncentráló kutatásokban hasznosul. Az utóbb említett lokális tempó pedig azokban a vizsgálatokban válik érdekessé, amelyek a tempó időbeli változásaira, dinamikai sajátossága-ira fókuszálnak. Az utóbbi megközelítéshez csatlakozva a disszertáció egyik fontos célja, hogy a beszéddallam és az intenzitás annotációjához alkalmazott modell keretében feldolgozhatóvá tegye a beszédtempó társas interakciókban tapasztalt modulációit is.

A beszédtempó számítógépes elemzéséhez nélkülözhetetlen kiindulópontot jelent, hogy a mérés alapját jelentő nyelvi egységek – amelyek időbeli sűrűségét vizsgálni kívánjuk – pozíciójáról információval rendelkezünk. A mérés ezen alapegységéül a szótagokat jelöltem ki, amelyek pozíciója és sűrűsége az alaphfrekvenciával és az intenzitásával ellentétben nem tartozik a közvetlenül mérhető akusztikai jellemzők közé. Létezik viszont egy Nivja de Jong és Ton Wempe által kidolgozott (de Jong & Wempe, 2009), a Praat program beépített szkript nyelvén implementált módszer, amely az intenzitásgörbe csúcsértékeinek vizsgálata útján egészen hatékonyan képes detektálni a beszédfolyam „zárványait” jelentő szótagmagokat.

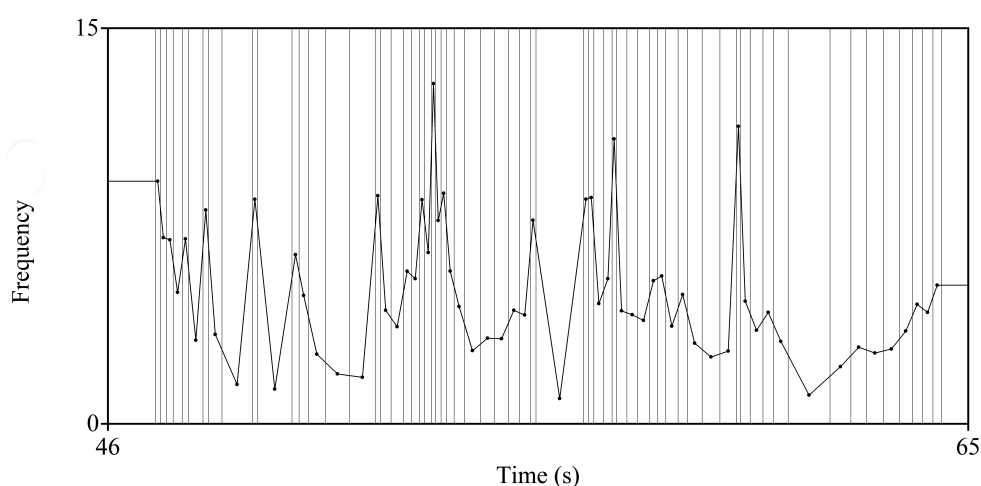


4.28. ábra. A szótagmagok pozíciójának detektálása (de Jong & Wempe, 2009).

A 4.28. ábra a program TextGrid formátumú kimenetét, a szótagmagok sorszámmal jelölt pozícióját az első annotációs szinten szemlélteti. A második szinten a vizsgált beszédrészlet fonetikus átírata, az annotációk fölött pedig az intenzitásgörbe látható. A szerzők által fejlesztett algoritmus első lépése az intenzitásgörbe csúcsértékeinek (az ábrán a fekete pontok jelzik

őket) kinyerése³⁵, amelyek nagy valószínűséggel jelzik a potenciális szótagmagok helyét. A következő lépés az intenzitás csúcsainak vizsgálata a csúcstarték és a következő csúcs közötti minimumérték különbségére vonatkozó küszöbérték (alapértelmezés szerint 2 decibel) segítségével az intenzitás kisebb kiugrásainak kiszűrése érdekében. Szintén eldobásra kerülnek a beszéd zöngétlen szegmentumaiban előforduló intenzitáscsúcsok (nagyobb energiával kiejtett zöngétlen mássalhangzók) is. Utóbbi esetre példa a 4.28. ábrán látható, a [ts] hang helyére eső 6. csúcs. Szintén kimarad a beszédrészlet elején, az első szótag előtt található csúcs (a felvételen itt beszédzaj hallható), mivel a közte és az őt követő csúcs közötti minimum, illetve az aktuális csúcs intenzitásértékének különbsége alatta van a szkript futtatása előtt beállított limitnek. Az algoritmus tökéletlen működésére példa a 7. szótag pozíciójának átcusúszása az [v] hangról a szintén zöngés [n] hangra a „tanul” szóban. A elcsúszás oka, hogy a szkript az intenzitáscsúcsok értékét mindig az őket követő csúccsal (illetve a kettő közötti minimummal) veti össze, ami alapján a később következő [n] helyett az [v] helyén található csúcspot dobja el. A hiba kiküszöbölése későbbi terveim között szerepel.

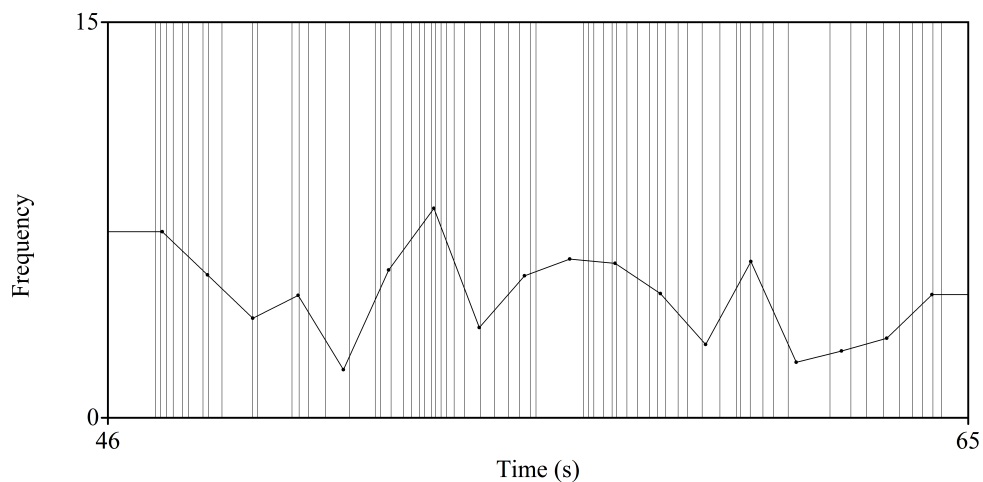
A szótagmagok detektálásának eredményét de Jong és Wempe a globális beszédtempó („speech rate”) és artikulációs tempó („articulation rate”) teljes hanganyagra számolt átlagának kiszámítására használták fel. Saját algoritmusomba egészen más célból, a beszédtempó lokális változásainak automatikus annotációja érdekében építettem bele, amelynek egyik köztes reprezentációs formája a szótagmagok pozíciója alapján megalkotott, a szótagok időbeli sűrűségének változását kifejező *szótagfrekvencia-görbe* volt.



4.29. ábra. A szótagok időbeli sűrűségének változása.

³⁵ de Jong és Wempe ehhez az intenzitásadatok tárolására szolgáló Praat objektumból („Intensity”) kiindulól, többlépcsős műveletsort („Down to Matrix”, „To Sound (slice)”, „To PointProcess (extrema)”) alkalmaz, de hasonló hatása van az előző szakaszban bemutatott „To IntensityTier (peaks)” parancsnak is.

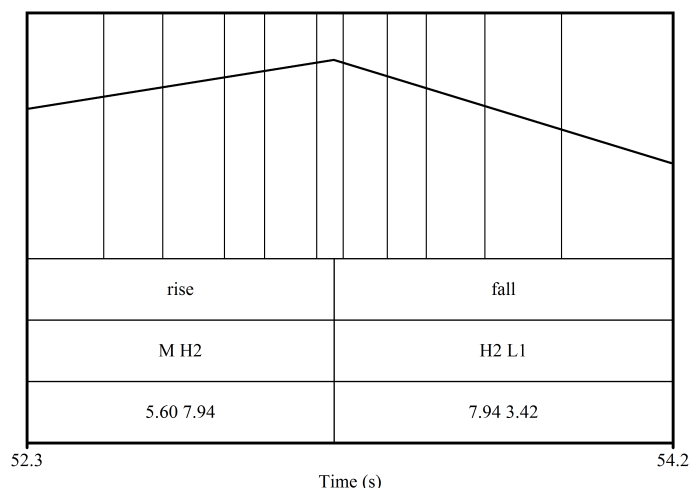
A 4.29. ábrán látható szótagfrekvencia-görbe egy adott pillanatban mérhető magassága a függőleges vonalakkal jelzett szótagmagok időbeli távolságával fordított arányban változik. A távolságok és a lokális szótagfrekvencia közötti viszony tehát a hanghullám periódusainak hossza és a hullám frekvenciája közötti viszony³⁶ analóg módon írható le. Az ábrán látható görbét a szótagmagok pozícióját tároló „TextGrid” objektumból konvertált „PointProcess” objektum alapján generáltam ugyanannak az algoritmusnak a használatával, amellyel a Praat programban a hanghullám periódushatárai alapján az alapfrekvencia leképezése is történik. A szótagmagok távolságának erős ingadozására való tekintettel a beszédtempó reprezentációja sokkal közelebb hozható a percepcióban megjelenő tempóélményhez, ha a szótagfrekvencia-görbén 1 másodperces időközönként mozgóátlagolást is végrehajtunk. Ennek eredményét szemlélteti a 4.30. ábra.



4.30. ábra. A lokális beszédtempó változását kifejező görbe.

A beszédtempó lokális változásainak szegmentálása és címkézése az átlagolt szótagfrekvencia-görbe, illetve a beszélőre jellemző globális tempó átlaga és szórása alapján, a beszéddallam és az intenzitás annotálásánál bemutatott módszerekkel történik. Az annotáció kimenetét a 4.31. ábra szemlélteti. A vertikális vonalak a szótagok pozícióját, a legalsó annotációs szinten lévő számok pedig a trendvonalak határain kapott sebességet (szótagszám / másodperc) mutatják.

³⁶ Minél hosszabb a periódusidő, annál alacsonyabb a frekvencia, illetve minél rövidebb a periódusidő, annál magasabb a frekvencia.



4.31. ábra. A beszédtempó annotációjának kimenete. A vertikális vonalak a szótagmagok pozícióját szemléltetik.

4.2.4. XML konverzió

A felhasználási lehetőségek kibővítése érdekében a ProsoTool aktuális verziója a Praat TextGrid formátumban tárolt címkék mellett XML kimenettel is kiegészült, amely a további fejlesztések során leginkább a stilizált dallamkontúrok web-alapú vizualizálásához bizonyult rendkívül előnyösnek. Mint az alább következő példában látható, a konverzió eredményeként kapott XML fájlokban az adatok rendezésének alapját a hangfelvételeken hallható beszélők jelentik, egy-egy <TIER> elem tartalmaként, egymástól elkülönítve tárolva a rájuk vonatkozó elemzéseket.

A <TIER> elemeken belül a <META> elem az adott beszélő hangterjedelmére vonatkozó globális információkat tartalmazza a beszélő azonosítójával (a példában: „Darinka”), átlagos alaphangfrekvenciájával, az alaphangfrekvencia-adatok szórásával, minimumával és maximumával, illetve az értékek eloszlása alapján számolt tartományok (L_1 , L_2 , M , H_1 , H_2) határaival. A <SEGMENTS> elem a <SEG> elemekben már a konkrét dallamkontúrok paramétereit rögzíti, a <BEGIN> és az <END> elemekkel jelölve meg időbeli pozíciójukat.

```
<?xml version='1.0' encoding='UTF-8'?>
<pitch>
  <file>angela.xml</file>
  <soundfile>angela.wav</soundfile>
  <tier>
    <meta>
      <speaker>Darinka</speaker>
      <meanf0>217.657</meanf0>
```

```

        <stdev>59.415</stdev>
        <divf0range1>158.242</divf0range1>
        <divf0range2>197.852</divf0range2>
        <divf0range3>237.462</divf0range3>
        <divf0range4>277.072</divf0range4>
        <minf0>133.768</minf0>
        <maxf0>369.865</maxf0>
    </meta>
    <segments>
        <seg>
            <label>ascending</label>
            <begin>0.060</begin>
            <end>0.164</end>
            <level1>H1</level1>
            <level2>H1</level2>
            <value1>249.240</value1>
            <value2>258.440</value2>
        </seg>
    </segments>

```

A közvetlenül a gyökérelem („PITCH”) alá tartozó <SOUNDFILE> elem a beszéddallam elemzését a megfelelő hangfájlhoz társítja hozzá, míg a <FILE> elem a beszélőváltást tartalmazó, szintén XML formátumú annotációra mutat rá, amelynek szerkezetét az alábbi példa szemlélteti.

```

<?xml version='1.0' encoding='UTF-8'?>
<?xml-stylesheet type='text/xsl' href='style_web.xsl'?>
<dialogue>
    <meta>
        <file>006mc22_F_shure</file>
        <soundfile>006mc22_F_shure.wav</soundfile>
    </meta>
    <sp>
        <speaker>agent</speaker>
        <speech begin='0.993' end='1.673'>tessek!</speech>
    </sp>
    <sp>
        <speaker>speaker</speaker>
    </sp>

```

```

        <speech begin='4.095' end='4.581'>jo napot</speech>
    </sp>
    <overlap>
        <sp>
            <speaker>speaker</speaker>
            <speech begin='4.581' end='5.079'>kivanok!</speech>
        </sp>
        <sp>
            <speaker>agent</speaker>
            <speech begin='4.581' end='5.079'>jo na</speech>
        </sp>
    </overlap>

```

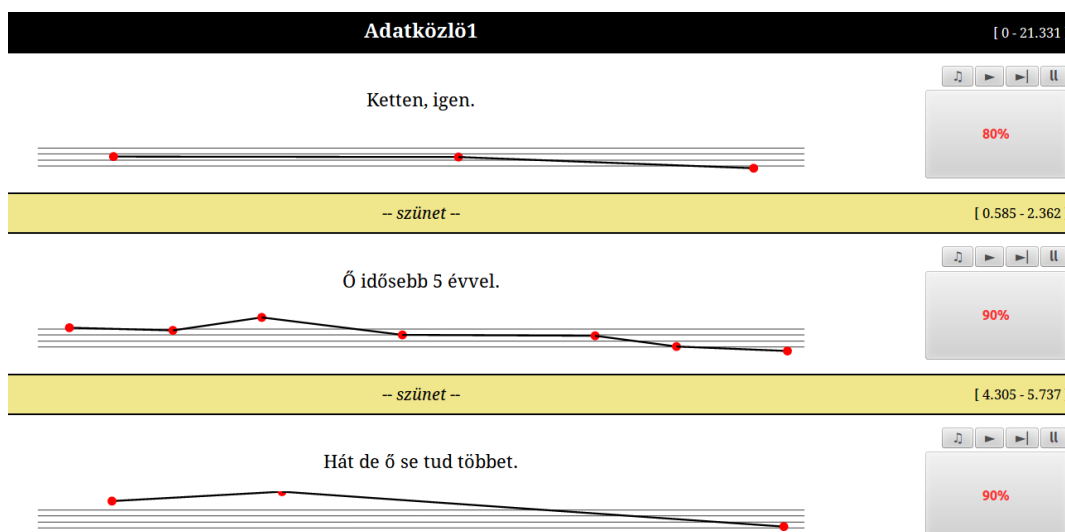
A fenti példában szereplő, szöveges transzkripcióval ellátott XML kód <SP> elemei az interakció egyes fordulóit tartalmazzák a beszélő nevével és az átírat szövegével, ahol a <SPEECH> elem attribútumai a kérdéses megnyilatkozás időbeli kezdetét és végét is megjelölik. Átfedő beszédek esetén az egyes beszélőkhöz tartozó részeket az <OVERLAP> elem fogja át.

4.2.5. A kimenet vizualizálása XSLT transzformációval

A szöveges transzkripciók és a beszéddallam elemzésének vizualizációja az XML fájlok XSLT transzformációja útján valósítható meg. A ProsoTool fejlesztése során ehhez kétféle XSL stíluslap készült el: egy az interakciók globális áttekintéséhez, egy másik pedig a beszéddallam részletesebb elemzéséhez. A stíluslapok elkészítését elsődlegesen az a szándék motiválta, hogy a kutatás eredményei ne csak a szűk szakmai, de a tágabb laikus közönség számára is hozzáférhetővé váljanak egy speciális eszközök nem igénylő, szemléletes és interaktív környezetben, amely – mint a 4.2.6. fejezetben látni fogjuk – a módszer szubjektív értékeléseken keresztül történő validációjához is hasznos eszköznek bizonyult.

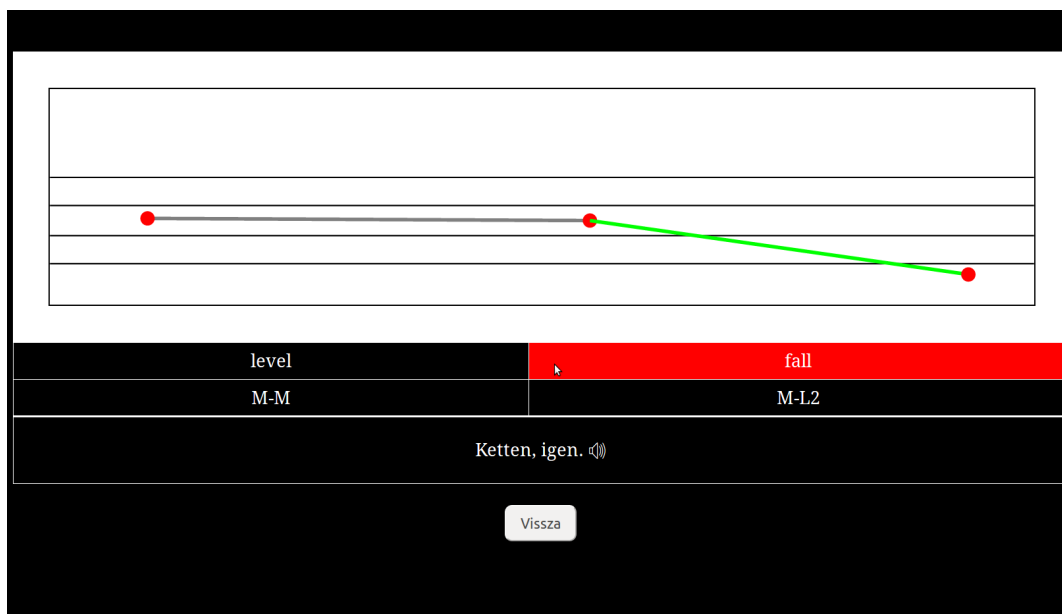
A megjelenítéshez használt XSLT transzformáció legfontosabb előnye, hogy az XML-ben tárolt adatok és a stíluslapok teljesen függetlenül kezelhetők egymástól, így a transzformáció a ProsoTool tetszőleges kimenetére alkalmazható. Maga az átalakítás kliensoldalon csupán egy webböngészőt igényel, amely az eredményül kapott XHTML kódot weblapként jeleníti meg. Az XML fájlok és a stíluslapok összekapcsolása történhet az XML dokumentumban elhelyezett hivatkozással (az előző szakaszban erre láttunk példát) vagy Javascript segítségével kliensoldalon, illetve PHP használatával a szerveroldalon történő transzformációhoz.

A fordulóváltásokat és a szöveges átiratokat tartalmazó annotációk XML formátumba konvertálását egy Praat szkript segítségével végeztem el. Ezek az annotációk képezik a globális áttekintéshez szolgáló, az 4.32 ábrán látható XHTML felület alapját, amely a szöveges átiratok alatt a ProsoTool kimeneteként kapott dallamkontúrok SVG grafikaként történő vizualizációjával is kiegészül.



4.32. ábra. A XSLT transzformáció eredményeként kapott interaktív felület.

Az interaktív felület jobb oldalán látható kezelőgombok segítségével lehetőség van az egyes megnyilatkozások meghallgatására, illetve a teljes interakció lejátszására is. Lejátszás közben a szöveges átiratok alatt látható dallamkontúrban az aktuális dallammenet színelváltozása mutatja, hogy éppen melyik szegmens hallható. Ezeknek az interaktív funkcióknak a vezérlése az XSL stíluslapban elhelyezett Javascriptekkel történik, amelyek a szükséges információkat (pl. lejátszandó megnyilatkozás időbeli kezdete és vége) az XML fájlokból olvassák ki.



4.33. ábra. A kimenet XSLT transzformációja interaktív SVG grafikaként.

A hangjegyeket ábrázoló kezelőgombra kattintva az egyes megnyilatkozások dallamszerkezetének a 4.33. ábrán látható, részletes elemzésére nyílik lehetőség. Az oldal felső részén elhelyezett SVG grafikát egy JavaScript kód generálja le a kérdéses beszélőre és megnyilatkozásra vonatkozó, XML fájlokból származó adatok alapján. Az ábra háttérét adó „négyvonalas kotta” a beszélő hangterjedelmének 5 részre felosztott tartományait szemlélteti, amelyekben az egyes dallammenetek egy kiinduló és záró hangmagasságot összekötő vektorként helyezkednek el. A dallammenetek alatt a ProsoTool kimenetében szereplő szöveges kategóriák (az ábrán: „level”, „fall”) és a kontúrok szerkezetét leíró relatív hangmagasságpárok (az ábrán: „M M”, „M L2”) mellett a megnyilatkozáshoz tartozó szöveges átírat is leolvasható. Az áttekintő felülethez hasonlóan itt is adott a lehetőség a megnyilatkozásokhoz és az egyes dallamszegmentumokhoz tartozó hangrészletek lejátszására, ami az ábrán látható kontúrokra vagy a szöveges blokkokra kattintva indítható el.

4.2.6. A kimenet értékelése

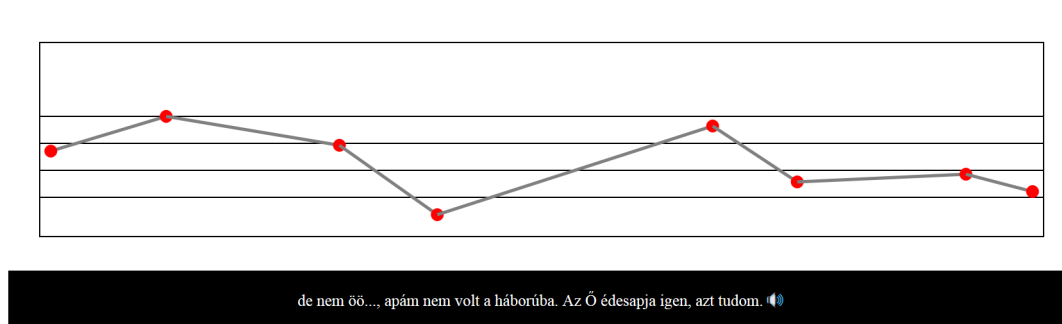
A vizualizációkat tartalmazó interaktív felület lehetőségeit kiaknázva később egy komplett weboldal³⁷ is készült a ProsoTool kimenetének szubjektív, percepcióhoz viszonyított értékelésére. Az értékelés két fordulóból állt. Az első fordulóban 25 megnyilatkozáshoz tartozó vizualizációt kellett az elemzés forrásául szolgáló hangfelvételekkel összevetni. A megnyilatkozások egy, a Debreceni Egyetem Néprajztudományi Tanszékének archívumából kölcsönzött interjú

³⁷ <http://lingua.arts.unideb.hu/pttest>

hanganyagából származó részlet (az adatközlő világháborús emlékeiről beszél) időben egymást követő közléseivel voltak azonosak, amelyek a teszt első fordulójában az elhangzás sorrendjében kerültek bemutatásra. A hangfelvétel az adatközlő lakhelyén, diktafon segítségével került rögzítésre egy gyűjtés során. A résztvevőknek 10-es skálán kellett meghatározni, hogy az egyes megnyilatkozásokhoz készült vizualizációk (lásd. 4.34. ábra) összbenyomásuk alapján mennyire feleltethetők meg a beszéd általuk érzékelt dallamával, más szóval, mennyire állnak kongruens viszonyban a látottak a hallottakkal:

„A következő oldalakon hosszabb-rövidebb beszédrészletek dallamának vizuális ábrázolását fogja látni egy «négyvonalas kottában». A feladat annak az értékelése lesz, hogy az ábrán látható dallam szubjektív benyomása alapján mennyire (hány %-ban) felel meg a felvételen hallhatónak.”

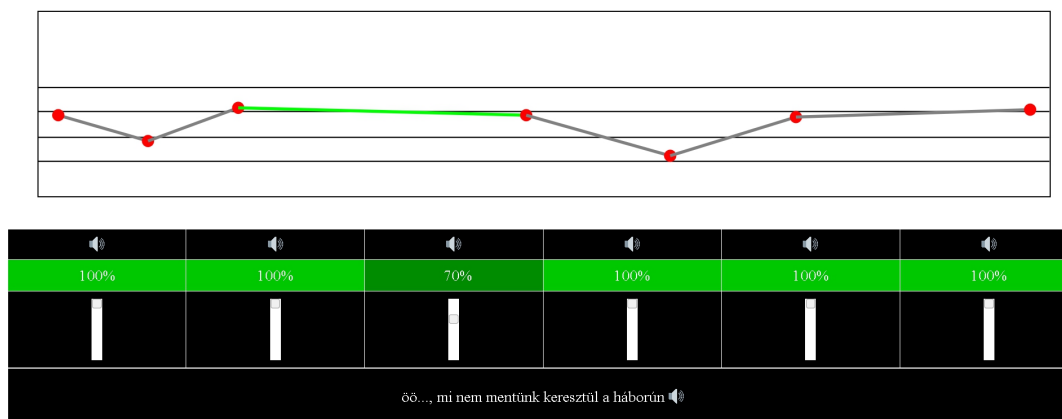
Az alkalmazott módszert a hallás és a látás modalitásainak a percepcióban való összekapcsolására vonatkozó korábbi kutatásaink (Hunyadi et al., 2014, 2015) motiválták, ami alapján várhattuk, hogy a két modalitás egy időben való értelmezése (kongruens vagy inkongruens), az automatikus annotáció vizuális kimenete könnyebben összevethető a hallható intonációval.



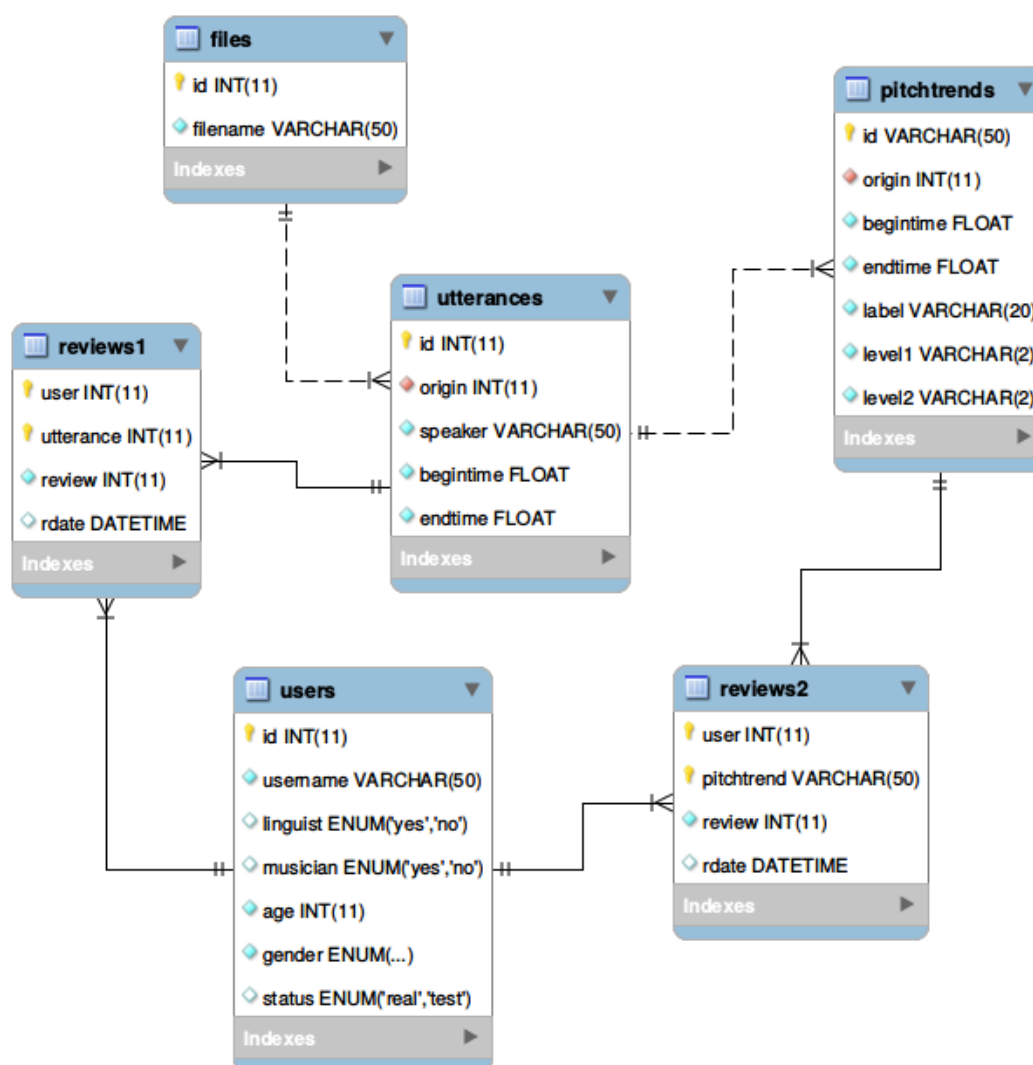
4.34. ábra. Az értékelőfelület első fordulójában látható vizualizációk.

A kísérlet második fordulójában ugyanezek a megnyilatkozásokat véletlenszerű sorrendben prezentáltam, az értékelőknek pedig dallammenetről dallammenetre volt lehetősége pontozni a vizualizációkat:

„A most következő oldalakon a korábban értékelt beszédrészleteket kell újra végighallgatnia véletlenszerű sorrendben. A feladat ezúttal is az ábrán látható és a felvételen hallható dallam összehasonlítása. A különbség, hogy most lehetősége van az elemzés részletes, dallammenetenként (a piros pontok által határolt vonalak) történő értékelésére.”



4.35. ábra. Szegmentumonként történő értékelés a 2. fordulóban. A értékelők munkáját egy Javascript animáció segítette, ami az aktuálisan értékelt szegmentum színének elváltoztatásával jelezte (lásd az ábrán balról a harmadik), hogy az adott pontszám éppen melyikre vonatkozik.



4.36. ábra. Az értékelő felület adatainak tárolására létrehozott MySQL adatbázis EER diagramja.

Az értékeléseket jelen fejezet írásáig 9 alany (nemek eloszlása: 3 férfi és 6 nő) végezte el, átlagéletkoruk 28 év, hárman rendelkeznek zenei háttérrel, négyen pedig nyelvészeti tanul-

4.2. táblázat. Az 10-es skálán kapott értékelések átlageredménye és szórása.

	Megnyilatkozások	Szegmentumok
Átlag	8.29	9.35
Szórás	2.05	1.51

mányokat folytatnak. Az alanyok kiválasztásánál ez utóbbi kettő (zenei és nyelvészeti háttér) nem jelentett szempontot, de a kérdőív elején fontosnak éreztem ezeket az adatokat is rögzíteni. A vonatkozó prozódiai jelenségekkel kapcsolatos képzésben a kísérletet megelőzően egyikőjük sem részesült. Az értékelések egy MySQL adatbázisban kerültek rögzítésre, az értékelők, a megnyilatkozások és a ProsoTool kimeneteként kapott dallamszegmentumok adataival együtt. A 4.36. ábrán az adatbázis modelljének EER-diagramja látható, ahol a *reviews1* relációs tábla tartalmazza az első, a *reviews2* pedig a második forduló értékeléseit, összekapcsolva azokat az illetékes résztvevővel (*users* tábla) és megnyilatkozással (*utterances* tábla), illetve a megnyilatkozásokhoz tartozó dallamszegmentumokkal (*pitchtrends* tábla). Az értékelés jelenlegi állását a 4.2. táblázat foglalja össze, de a megfelelő SQL lekérdezéssel a pontosabb részletek is megvilágíthatók:

```
SELECT reviewed.utt,reviewed.duration,
GROUP_CONCAT(pitchtrends.label SEPARATOR ',') AS contours,
reviewed.review1_avg, reviewed.review2_avg, reviewed.erate
FROM (SELECT utterances.id AS utt,
        ROUND(utterances.endtime - utterances.begintime,2) AS duration,
        ROUND(AVG(reviews1.review),2) AS review1_avg,
        ROUND(AVG(reviews2.review),2) AS review2_avg,
        ROUND((COUNT(CASE WHEN reviews2.review < 10 THEN
reviews2.review END) / COUNT(reviews2.review)) * 100,2)
        AS erate
FROM utterances INNER JOIN reviews1
ON utterances.id = reviews1.utterance
INNER JOIN pitchtrends
ON utterances.id = pitchtrends.origin
INNER JOIN reviews2
ON pitchtrends.id = reviews2.pitchtrend
GROUP BY utterances.id) AS reviewed
```

INNER JOIN pitchtrends ON reviewed.utt = pitchtrends.origin

GROUP BY reviewed.utt

A lekérdezés eredménye a 4.3. táblázatban látható, ahol a harmadik oszlop az egyes megnyilatkozásokra kapott kimenetet (egy-egy betűjellel címkézett dallamszegmentumok sorozata), a harmadik és negyedik oszlop pedig a két fordulóban kapott értékelések átlageredményét tartalmazza. Az utolsó oszlopban szereplő hibaszázalékok a megnyilatkozáshoz tartozó szegmentumokra kapott azon értékelések arányát mutatják, amelyek valamilyen problémára (inkongruenciára) utalva az alapértelmezés szerinti maximumnál kisebb pontszámot rögzítenek.

4.3. táblázat. Az egyes megnyilatkozások elemzésére kapott globális (1. forduló) és szegmentumonkénti (2. forduló) értékelések átlageredményei. A dallamszegmentumokra vonatkozó rövidítések feloldása: (R)ise, (F)all, (A)scending, (D)escending, (L)evel.

ID	Időtartam	Dallamszegmentumok	1. Ford.	2. Ford.	Hibaaarány(%)
1	0.58	L F	8.00	8.61	33.33
2	1.94	F D R F L F	8.67	9.52	22.22
3	1.61	R F	8.89	9.56	11.11
4	2.28	F R F	7.67	9.41	33.33
5	2.53	D F	8.44	9.33	22.22
6	1.26	F R A	9.11	9.44	18.52
7	3.29	R F R F F L	9.00	9.13	29.63
12	2.11	F R D F R L	8.67	9.50	18.52
13	1.60	A	9.00	8.22	55.56
14	0.84	F R	8.67	9.00	22.22
15	1.67	R F R F	8.56	9.53	13.89
17	4.23	R F L F R R F R F R F	9.11	9.29	16.16
20	0.32	R	7.44	8.11	22.22
21	1.76	D F R D	8.33	9.42	25.00
22	1.01	F L	7.89	9.28	16.67
24	1.62	R D F L	7.78	9.22	27.78
25	5.26	A R L F L R F F A R F D R A F	7.89	9.68	11.85
26	1.41	L	7.67	8.67	33.33
27	1.51	L	7.56	8.44	66.67
29	1.49	R F	9.11	9.78	11.11
30	1.60	R F	7.22	9.11	27.78
31	5.29	R L F A F R F R F R D	8.00	9.40	25.25
32	8.20	A F R F L R R D R L F R L F F R	7.56	9.39	18.06
34	2.93	R F R F R F	8.11	9.26	35.19
35	1.11	A F	9.00	9.11	22.22

A táblázatból jól látszik, hogy a második fordulóból származó értékelések átlaga minden esetben meghaladja az első fordulót. Ebben nagy valószínűséggel szerepet játszik az a nem elhanyagolható tényező, hogy a második fordulóban csak azoknál a dallamszegmentumoknál történt pontozás, ahol az értékelő inkongruensnek találta az elemzés kimenetét a beszéddallam

általára érzékelt modulációjával. A többi szegmentum esetén a maximális pontszám nyugtázásával jelezte egyetértését, így a szegmentumokra kapott értékelések átlaga könnyen meghaladhatta a megnyilatkozás egészére adott pontszámot, ha az összbenyomást egyetlen hibás dallammenet vagy az elemzés túlzott részletessége (lásd 25. és 32. megnyilatkozás) rontotta le. Ezen – az eredményeket pozitívan torzítható – tényező kompenzálására történt a táblázat utolsó oszlopában szereplő, a pontszámokat bináris formában (van jelzés, nincs jelzés) feldolgozó hibaarányok mérése, amelyek összes megnyilatkozásra számolt átlaga 25.59 százalék. A megnyilatkozások egészére adott értékelések a megnyilatkozás hosszával csak gyenge korrelációt ($R = -0.01$) mutattak és az egyes alanyok véleménye között sem találtam szignifikáns különbséget:

$$\chi^2(114, N = 9) = 138, p = 0.062 \quad (4.8)$$

Az alábbi lekérdezés a különböző típusú dallamkontúrokra adott értékelések számáról, illetve az inkongruenciára utaló jelzések arányáról informál (a lekérdezés eredményei a 4.4. táblázatban tekinthetők meg):

```
SELECT pitchtrends.label,
COUNT(allrev.review) AS allint2,
COUNT(CASE WHEN allrev.review < 10 THEN allrev.review END) AS wrong,
ROUND((COUNT(CASE WHEN allrev.review < 10 THEN allrev.review END)
/ COUNT(allrev.review)) * 100,2) AS wrongrate FROM pitchtrends
INNER JOIN (select * from reviews2)
AS allrev ON pitchtrends.id = allrev.pitchtrend
GROUP BY pitchtrends.label
```

4.4. táblázat. A különböző típusú dallamkontúrokra adott értékelések.

Kontúr	Értékelés	Inkongruens	Hibaarány(%)
ascending	72	15	20.83
descending	81	20	24.69
fall	405	91	22.47
level	135	32	23.70
rise	342	66	19.30

Habár az *eső* („fall”) dallamkontúr uralkodó előfordulásából kifolyólag az inkongruens esetek nagy százaléka (40.63%) is ehhez a típushoz köthető, a hibaarány tekintetében nem tapasztalható jelentős különbség a kontúrok között. A dobogós helyezést kapó *ereszkedő* („descending”) és *színttartó* („level”) kontúrok inkongruenciája a legtöbb esetben valószínűleg annak

köszönhető, hogy a túlságosan erős stilizáció hallható modulációkat fedett el a beszéd eredeti dallamszerkezetében (példaként lásd a 4.3. táblázat 27. és 28. megnyilatkozását egyetlen kontúrban egyesítő kimenetet).

5. FEJEZET

Percepciók kísérletek

Az alábbi fejezet a disszertáció alaphipotézisének különböző aspektusokból történő körbejárására és igazolása kidolgozott percepciók kísérletek leírását tartalmazza. A kísérletek közös célja annak a vizsgálata volt, hogy a verbális tartalom hiányában, kizárólag egy társalgás vagy egy izolált megnyilatkozás prozódiai jellemzőire hagyatkozva milyen, a lehetséges kontextusra, az interakcióban résztvevők viszonyára vagy a tematikus szerkezetre vonatkozó következtetéseket tudunk levonni. A második kísérlet annyiból kivétel, hogy ott nem prozódian túlmutató információk feltárása, hanem a jellemzők időbeli változásának („emelkedik vagy csökken?”) a követése volt a feladat. A tesztek lebonyolításánál változatos kondíciókat és kísérleti módszereket alkalmaztam. A verbális tartalom kizárását az első két kísérletben az alanyok számára ismeretlen nyelv felhasználásával, illetve újraszintetizált, a beszéd szegmentális elemeit rejtve hagyó felvételekkel biztosítottam. A harmadik kísérletnél („Angéla néni esete”) a verbális tartalom tekintetében változatlan, csak a prozódia és a közlés attitűdtartalma szempontjából modulált megnyilatkozások (és azok újraszintetizált verziói) segítségével értem el ugyanezt. A társalgási szituáció beazonosítását célzó („formális vagy informális”) utolsó kísérletet azonos kondíciók mellett tanított neurális hálózatokkal is megismételtük, aminek eredményei a 6. fejezet első részében találhatók meg.

5.1. Idegen fülek a Perzsa öbölben

Az elsőként bemutatásra kerülő kísérlet különböző forrásból származó dialógusrészletek globális jellemzőinek megítélését vonta vizsgálat alá olyan „unimodális” kondíciók között, amelyben a vizualitás, a megnyilatkozások szemantikai tartalma és a kontextusra vonatkozó háttérismerek kizárásával a megfigyelők válaszadás közben – egyéb elérhető információ hiányában – csak az interakció prozódiai szerkezetére hagyatkozhattak.

A kísérleti alanyoknak egy általuk nem beszélt idegen nyelven folyó dialógusok hangfelvételeiből kivágott részleteket kellett végighallgatniuk, majd az adott dialógussal kapcsolatos tesztkérdésekre választ adniuk. A kérdések a dialógusok főntebb említett globális jellemzőire, a résztvevők viszonyára, a kivágott részlet a dialógus egészében elfoglalt pozíciójára és a benne észlelhető eseményekre irányultak. Az ismeretlen nyelv használatának célja az volt, hogy a kérdések megválaszolása közben az alanyok számára a beszélgetés verbális tartalma a szemantika tekintetében ne legyen hozzáférhető, csakis a társalgás prosódiai jegyeire tudjanak hagyatkozni.

A kísérletek irányított körülmények között, csoportos részvétellel zajlottak. A résztvevők (összesen: 37 fő és 4 kísérleti alkalom, résztvevők átlagéletkora: 20 év) közösen hallgatták végig a felvételeket, majd egy e célra készített webes felület¹ segítségével válaszoltak a felvételekkel kapcsolatos kérdésekre, amelyeket szöveges indoklással is el kellett látniuk. A felvételek lejátszására tantermi körülmények között került sor, hangszóró használatával. A lejátszások minden válaszadás előtt megismétlődtek, és mivel minden felvétel esetében azonos kérdésekre kellett választ adni, az alanyok előre ismerhették ezeket.² A többszöri meghallgatást az motíválta, hogy az alanyoknak lehetősége legyen mindig csak az aktuális megfigyelési szempontra koncentrálni.

5.1.1. A kísérleti felvételek

A kísérletben összesen négy hangfelvétel szerepelt, amelyek különböző perzsa nyelven folyó, körülbelül 15-20 perc hosszúságú dialógusok egy perces részleteit tartalmazták. Az eredeti dialógusok a HuComTech kutatócsoport laboratóriumában kerültek rögzítésre egy hangszigetelt felvevő helyiség és két AT2035 típusú, kardioid iránykarakterisztikájú, állványon rögzített stúdió mikrofon használatával. A társalgás során a résztvevők ülő helyzetben kommunikáltak egymással, a rájuk irányzott felvevőeszközöktől végig azonos távolságra. A felvételeknél használt beállításokat az 5.1. ábra szemlélteti. A hanganyag két külön csatornán tartalmazza a dialógusban részt vevők beszédét, ezek automatikus szétválasztása azonban a mikrofonok áthallása miatt jelenleg nem megoldott. A rögzített csatornák jel-zaj viszonyaival kapcsolatos adatok az 5.1. táblázatban láthatók.

A felvételek mindegyikén perzsa anyanyelvű résztvevők szerepelnek. A dialógusok egy része előzetes instrukciók szerint, előre kiválasztott témák mentén zajlott, amelyeket az egyik be-

¹ <http://lingua.arts.unideb.hu/szekrenyes/experiments/persian/test-v1/>

² A kérdések a tesztfelületen sem különültek el, ugyanazon oldalon egyszerre láthatók voltak. A kutatás további tervei között szerepel a kísérlet eltérő kondíciók között való megismétlése: ahol minden alany csak egy felvételhez kötődően kell válaszokat adnia, és nem ismeri előre a kísérleti kérdéseket sem.



5.1. ábra. Felvételi beállítások.

5.1. táblázat. A felvételeken hossza (másodpercben) és SNR értéke (a beszéd- és szünetminták RMS értékei alapján számolt jel-zaj viszony decibelben kifejezve).

Felvétel	Hossz	SNR
1	50.05s	23.2 db
2	60.86s	25 db
3	61.18s	23.4 db
4	55.6	27.8 db

szélgető fél interjúvezetőként vezetett fel a társalgás során. A felvezetés általában olyan kérdések (pl. „Beszélne nekem egy pozitív élményről, amely mostanában ért?”) segítségével történt, amelyek a HuComTech korpusz építése során használt mintát (Pápay et al., 2011) követve egy személyes hangvételű, a kísérleti körülményekhez képest természetes és közvetlen hangulatú kommunikációs légkör megteremtését célozták. A dialógusok másik része előzetes instrukciók nélkül, de egy beépített, a beszélgetés fenntartásáért felelős irányító személy részvételével bonyolódott le. A kivágott részleteken hallható beszélgetések atmoszféráját erősen befolyásolta az éppen érintett téma és a beszélgető felek viszonya is. A kísérletben használt felvételrészletek jellemzőinek összegezése az 5.2. táblázatban olvasható. A „Dialógus menete” oszlopban szereplő értékek saját szubjektív benyomásaimat rögzítik.

5.2. táblázat. A felvételek jellemzőinek összefoglaló táblázata.

Sorsz.	Jelleg	Irányító fél	Beszélők viszonya	Dialógus menete	Témaváltás	Részlet pozíciója
1	irányított	nő	ismerős	vontatott	motiválatlan	közepén
2	kötetlen	férfi	ismerős	könnyed	motivált	elején
3	kötetlen	férfi	idegen	vontatott	motivált	végén
4	irányított	nő	ismerős	pergő	nincs	közepén

Az 1-es sorszámú felvétel egy interjúbeszélgetés részletét tartalmazza, amely egy témaváltás határánál, a dialógus közepén helyezkedik el. A férfi résztvevő a részlet első felében éppen élménybeszámolójának végére ér, amikor a beszélgetést kezdeményező nő az előzetes instrukciókhoz tartva magát lezárja a témát, majd megkéri, hogy térjenek át egy másikra.

A 2-es sorszámú felvételen két közeli ismerős társalgását halljuk egészen a dialógus elejétől. A párbeszéd vidám és oldott hangulatban, személyes jellegű kérdésekkel indul, majd egy általánosabb témájú, de az előzményekhez szorosan köthető irányában folytatódik.³

A 3-as sorszámú felvétel esetében egy éppen kimerülőben lévő társalgási téma végénél kapcsolódunk bele a dialógusba, ahol a beszélgetést irányító férfi egy hirtelen kérdéssel megváltoztatja a beszélgetés irányát, amelybe egyre aktívabban kapcsolódik bele a női fél, akinek szerepe a felvétel elején csupán rövid visszacsatolásokra korlátozódott. A férfi beszélő személye megegyezik az előző (2-es) felvételen hallhatóval.

A 4-es sorszámú felvételen egy interjúbeszélgetés részlete hallható, ahol éppen egy különösen személyes téma körül forog a beszélgetés, amely a részlet időtartama alatt nem változik meg. A felek ismerik egymást, aktívan és határozott állásponttal vesznek részt a dialógusban, gyakran egymás szavába vágnek. A női beszélő személye megegyezik az 1-es felvételen hallhatóval.

A 5.3. táblázatban a felvételek azon temporális jellemzői szerepelnek (pl. szünetek és átfedő beszédek időbeli hossza), amelyek viszonyai – hipotézisem szerint – a kísérleti kérdések megválaszolása közben hatással lehettek az alanyok döntéseire.

Az adatok alapján három megkülönböztető tényező figyelhető meg:

1. Az első és a negyedik felvételnél – ahol az interjúvezető szerepét ugyanaz a hölgy töltötte be – a résztvevők megszólalásainak aránya kevésbé kiegyenlített (a férfi többet beszél).

³ A vizsgaidőszak aktuális terhei után arról kezdenek társalogni, hogy „mindennek megvan a maga ideje az ember életében, így a tanulásnak is.

5.3. táblázat. A felvételek temporális viszonyai másodpercben.

Sorsz.	Férfi beszél	Nő beszél	Szünetek	Átfedő beszéd
1	28.24	16.48	3.85	1.10
2	27.33	30.63	2.14	0.00
3	30.51	24.47	2.03	0.13
4	26.97	17.85	2.62	7.39

2. Az első felvételen a többihez képest több szünet mérhető.
3. A negyedik felvételen jelentősen több átfedő beszéd mérhető.

5.1.2. A kísérleti kérdések

A kísérletben részt vevő alanyoknak minden felvétel esetében az alábbi 4 tesztkérdésre kellett választ adniuk:

1. Kérjük, benyomásai alapján döntse el, hogy a felvételen hallható beszélgetést jellegét tekintve hova sorolná? (Irányított / Kötetlen / Nem tudom eldönteni)
2. A beszélgető felek közül ki irányítja inkább a beszélgetés folyamatát? (A férfi beszélő / A női beszélő / Nem tudom eldönteni)
3. Hova helyezné el a felvételen hallható részletet a beszélgetés egészében? Valahol a(z)... (Előjén / Közepén / Vége felé / Nem tudom eldönteni)
4. Van-e olyan benyomása a felvétel lejátszása során, hogy a felvételen hallható beszélgetés témája megváltozik? (Van / Nincs / Nem tudom eldönteni)

Az első kérdés célja annak felmérése volt, hogy a kísérleti alanyokban alakult-e ki valamilyen összbenyomás a felvétel meghallgatása után, képesek-e globálisan értékelni az interakciót a kontextus és a verbális tartalom ismerete nélkül. A feladatot és az eredmények értelmezését több tényező is nehezítette. Egyrészt a két kategória (irányított, kötetlen) elmosódott, relativizálható határai és tág értelmezési lehetőségei, másrészt a dialógusrészletek 5.2. táblázatban megadott (az alanyok számára ismeretlen) paramétereik között több olyan jellemző is szerepel (pl. a résztvevők viszonya), amely szintén befolyásolhatta a társalgás hovatartozásának megítélését. A független változók hasonló kölcsönhatásaival a többi kérdés esetében is számolni kellett. A domináns beszélő kiválasztásánál például meghatározó tényező lehet a dialógus jellege: egy kötetlen párbeszéd esetén valószínűleg nehezebben eldönthető, ki az irányító fél. A dialógus kötetlenné tűnő atmoszférája ugyanakkor – az interakció egészében elfoglalt pozíció megítélésére hatva – utalhat a beszélgetés előre haladott állapotára is („a kellős közepén jár”),

míg a vontatottnak, erőltetettnek (irányítottak) tűnő interakció a beszélgetés közeledő végét jelezheti, vagy éppen azt, hogy csak most kezdődött („még nem melegedtek bele”).

Hipotézisem szerint a témaváltás esetében fontos tényező lehet a témaváltás motiváltsága. Ha egy témaváltás motivált, más szóval a két téma indokolható kapcsolatban, netalán hierarchikus viszonyban (téma és altéma) áll, akkor a téma megváltozása valószínűleg nehezebben észlelhető. A váratlanul jövő, motiválatlan témaváltások ellenben markánsabb nyomott hagy-
nak a társalgás dinamikai struktúrájában, megkönnyítve így az észlelést.

5.1.3. Eredmények

Az 5.4. táblázat az első kérdésre adott válaszok eloszlását mutatja. A 2. felvétel esetében a kérdezettek döntő többsége (91%) ítélte a dialógust kötetlennek, amit a felvétel paraméterei is megalapozottá tesznek. A többi esetben már jóval kevésbé egyértelmű (56–67%) a többség, ami valószínűleg az egymással ellentétes irányban ható független változók (pl ismerősök irányított keretek között) számlájára írandó. A harmadik és a negyedik felvételnél a válaszok többsége ellentmondani látszik a társalgás tényleges kereteinek. Az előbbi esetében ez a résztvevők viszonyával („idegenek”) és a társalgás vontatottságával, az utóbbinál pedig a téma személyes jellegéből fakadó heves, sok átfedő beszéddel járó interakcióval magyarázható.

5.4. táblázat. Az első kérdésre (kötetlen vagy irányított) adott válaszok eloszlása.

Felvétel	Irányított	Kötetlen	Nem eldönthető
1	21	15	0
2	2	34	0
3	25	7	4
4	9	25	2

Mint az 5.5. táblázatban látható, az irányító fél tekintetében a válaszok sokkal megosztottabbak voltak. A „nem eldönthető” opció választása viszont az előző szakaszban említett vára-
kozásoknak megfelelően leginkább azoknál a felvételeknél jelentkezett, amelyet az előző kérdés során a többség kötetlennek ítélt.

5.5. táblázat. A második kérdésre (irányító fél) adott válaszok eloszlása.

Felvétel	Férfi	Nő	Nem eldönthető
1	19	17	0
2	14	6	16
3	15	16	5
4	16	9	11

Az első felvétel kivételével a válaszok a harmadik kérdés esetében is nagyon megosztottak (lásd 5.6. táblázat). A második felvétel esetében viszont elmondható, hogy a megkérdezetteknek

csak a 11 százaléka gondolta, hogy valahol a dialógus végén található. A negyedik felvételnél hasonlóképp helyesen zárkóztak el az interakció elejétől.

5.6. táblázat. A harmadik kérdésre (részlet pozíciója) adott válaszok eloszlása.

Felvétel	Elején	Közepén	Végén	Nem eldönthető
1	6	25	2	3
2	14	15	4	3
3	10	9	12	5
4	4	13	16	3

Ha csak a motiválatlan témaváltásokat számítjuk (1-es felvétel), a negyedik kérdés esetében a többségi döntés (54-62%) összhangban áll a felvételek tartalmával, de a felhasznált anyag kis mennyisége és az alanyokat potenciálisan befolyásoló tényezők sokasága miatt túlságosan messzemenő következtetéseket itt sem érdemes levonni.

5.7. táblázat. A negyedik kérdésre (témaváltás) adott válaszok eloszlása.

Felvétel	Van	Nincs	Nem eldönthető
1	21	14	1
2	12	22	2
3	9	20	7
4	9	23	4

5.2. Emelkedik vagy csökken ?

Az alábbi kísérletben három prozódiai jellemző, a hangmagasság, az intenzitás és beszédtempó változásainak előbeszédben való követhetősége került vizsgálat alá, amelyhez stimulusként ismét az előző, 5.1. szakaszban már bemutatott perzsa nyelvű hangfelvételeket használtam fel. Az idegen nyelvet itt is a szemantikai kizárása indokolta abból a célból, hogy az alanyok kizárólag akusztikai jellemzőkre koncentrálnak, nyelvi információk által nem terhelve reagáljanak a felvételekre. A tesztek lebonyolítása és kiértékelése során figyelembe kellett vennem a percepció azon speciális, korábbi munkákban (Tarnóczy, 1984; Hamill & Price, 2008) leírt sajátosságait is, miszerint az egyes prozódiai jellemzőkhöz kötődő hangérzetet egyszerre több fizikai paraméter is befolyásolhatja.⁴ Ezeket a fizikai változókat a hanganyag különböző módszerekkel szétválasztott szupraszegmentális vetületeinek leképezésével és újraszintetizálásával igyekeztem kontrollálni (lásd 5.2.2).

⁴ Két azonos amplitúdóval rendelkező szinuszhullám által kiváltott hangosságérzet például a hullámok frekvenciája függvényében is változik (Fechner, 1860).

5.2.1. Mérési módszerek

A kísérlet több fordulóból állt, amelyek során a négy felvétel eltérő variációit fejhallgatóval hallgatva kellett az alanyoknak valós időben jelezniük a három prozódiai jellemző valamelyikének változásait. A reakciók egy webes interfész⁵ segítségével kerültek rögzítésre, amelyen keresztül kétféle jelzés („növekszik” vagy „csökken”) volt leadható a billentyűzet megfelelő iránygombjainak lenyomásával, amikor az adott jellemző az alany számára érzékelhetően pozitív (pl. nő a beszédtempó) vagy negatív (pl. csökken a beszédtempó) irányban megváltozott. A gombnyomásokat mindig vizuális visszacsatolás követte egy 5 szintű (a felvétel elején a 3. szint jelentette a kiinduló helyzetet) állapotjelzőn⁶ mutatva a kérdéses prozódiai jellemző aktuális, az alany reakciót tükröző szintértékét.

Minden reakció esetén két attribútum került rögzítésre: a jelzés ideje (a lejátszás indítása után eltelt idő) és iránya (pozitív vagy negatív). A teljes hangfelvétel lejátszása után a felvételre kapott reakciók sorozata egy Praat TextGrid formátumú adatállományban tárolódott el, de az alanyok jelzései által vezérelt vizualizáció PHP fájlokból is visszajátszható egy e célból készült felületen.⁷ Mint az 5.2. ábrán látható, az előbbi formátum egyik előnye, hogy az adatok feldolgozásánál a Praat programmal egyszerűen visszakövethető, hogy az egyes jelzések a felvétel mely részleteinél történtek, habár az egyénenként és jelzásként változó reakcióidő miatt nem tudhatjuk biztosan, hogy egy bizonyos jelzést pontosan melyik akusztikai esemény váltotta ki.

5.2.2. Kísérleti feladatok és kondíciók

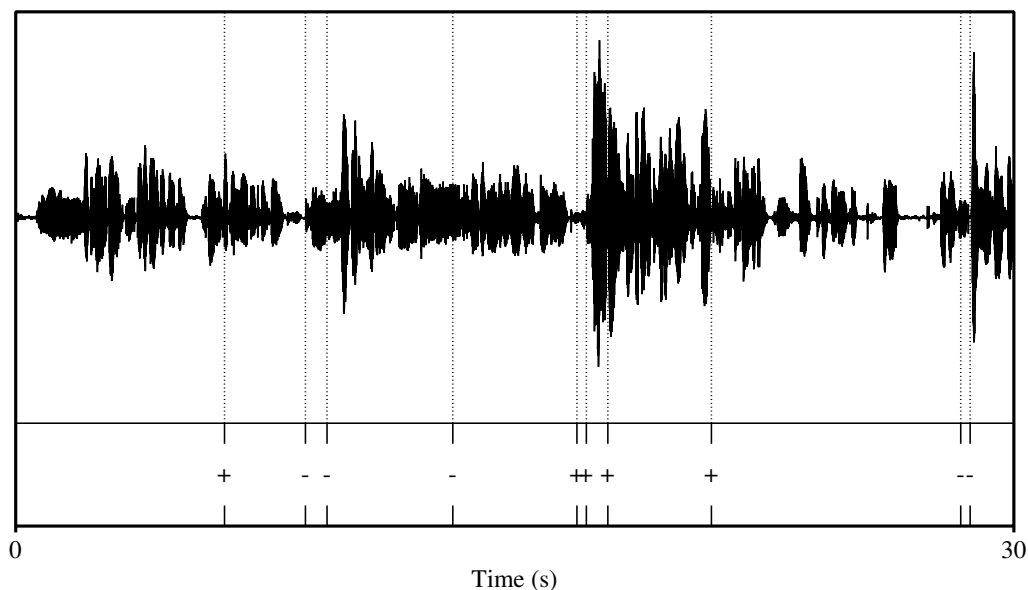
A kísérleti alanyok mind a négy dialógusrészletet ötféle variációban hallgatták végig, ahol az első variáció az eredeti hanganyaggal volt azonos, három az egyes prozódiai jellemzők izolált, egy pedig ugyanezek együttes, de az egyéb pszichofizikailag befolyással bíró tényezőktől (pl. a beszélők hangszíne, beszédsegmentumok) „megtisztított” reprezentációjára szolgált. A felvételek különböző módszerekkel történő konvertálása után tehát az alábbi öt verzió állt rendelkezésre:

1. Az eredeti hangfelvétel bármiféle manipuláció nélkül.
2. Az eredeti felvételből mért alapfrekvencia-adatok alapján generált, változó frekvenciájú gégehang. Az intonáció izolált reprezentációjára szolgált.

⁵ <http://lingua.arts.unideb.hu/perctest/index.html>

⁶ A kijelző egy több részre felosztott négyszöggént volt megjelenítve. A négyszög vertikális irányú, a szélső értékeket színváltozással is hangsúlyozó kitöltöttsége jelezte pl. a beszélő aktuális hangmagasságát.

⁷ <http://lingua.arts.unideb.hu/perctest/view.php>



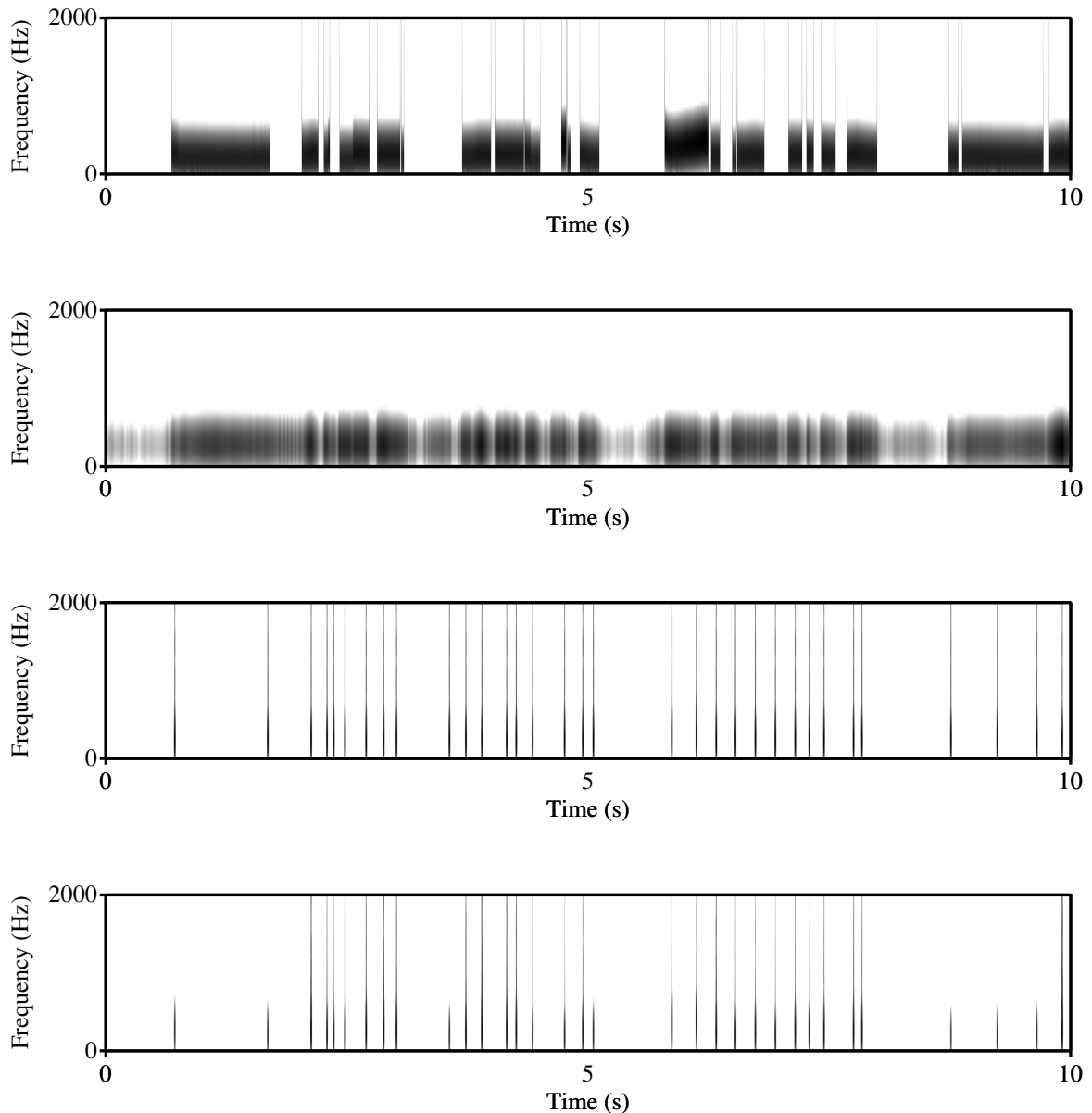
5.2. ábra. A prosódiai jellemzők változásaira adott reakciók (növekszik: „+”, csökken: „-”) ábrázolása a stimulusként szolgáló hangfelvétel időtengelyén.

3. Az eredeti felvételből mért intenzitásadatok alapján generált, változó amplitúdójú szinuszhullám. A beszédintenzitás izolált reprezentációjára szolgált.
4. Azonos intenzitású és frekvenciájú szinuszhangokból (rövid kattanások) álló impulzus-sorozat, amelyben minden hang az eredeti felvételen detektált szótagmagok helyét jelzi. A beszédtempó izolált reprezentációjára szolgált.
5. Változó intenzitású és frekvenciájú szinuszhangokból (rövid kattanások) álló impulzus-sorozat, amelyben minden hang az eredeti felvételen detektált szótagmagok helyét jelzi. A beszédtempó, az intonáció és a beszédintenzitás együttes reprezentációjára szolgált.

A konvertálást igénylő variációk közül a második előállítás a Praat program beépített eljárásaival egyszerűen, két lépésben („To Pitch...”, „To Sound (hum)”) kivitelezhető. A harmadik variáció létrehozásához első körben egy az eredeti felvétel időtartamával egyező szinuszhullám generálására volt szükség, amelynek intenzitás szerkezetét az eredetiből kimért intenzitásadatoknak megfelelően szintén a Praat programmal moduláltam. A beszédtempó reprezentációjául szolgáló negyedik és ötödik variáció egy komplikáltabb, a Praat program szkript-nyelvén implementált algoritmus⁸ kidolgozását igényelte, amelynek kimenetében a szótagmagok helyén felhangzó impulzusok (20 milliszekundum hosszú kattanások) sűrűsödése és ritkulása fejezte ki a beszédtempó változásait. A felvételek szintetizálásánál a beszélőváltás kézi annotációja biz-

⁸ A szótagmagok előzetes detektálását de Jong & Wempe (2009), a 4.2.3. fejezetben bemutatott eljárásával végeztem el.

tosította, hogy a különböző beszélőktől származó impulzusok külön csatornára (bal vagy jobb) kerüljenek. A 5.3. ábra az 1. felvétel alapján szintetizált variációk (2-5.) első 10 másodpercének spektrografikus ábrázolását jeleníti meg.



5.3. ábra. Az 1. felvétel átkonvertált variációnak (fentről lefelé: 1-5.) spektrografikus ábrázolása.

A kísérlet 9 fordulójában végrehajtandó feladatokat (melyik prosódiai jellemzőre kell koncentrálni és reagálni) és az egyes fordulókban használt, eltérő kondíciókat biztosító felvétel-verziókat az 5.8. táblázat foglalja össze. Mint a táblázatban látható, a reakciók mérése minden prosódiai jellemző esetében 3 fordulóból állt, ahol az első kondíciót a felvétel eredeti, a beszélő valódi hangját rögzítő verziójának meghallgatása jelentette. Ezután következett a kérdéses jellemző leképezése útján újraszintetizált variáció, amelyben a többi fizikai paraméter hatása

5.8. táblázat. A kísérlet 9 fordulójában különböző kondíciók között végrehajtandó feladatok.

Forduló	Feladat	Variáció	Beszéd	Hangmagasság	Intenzitás	Tempó
1	Tempó	1	X	X	X	X
2	Tempó	4	-	-	-	X
3	Tempó	5	-	X	X	X
4	Dallam	1	X	X	X	X
5	Dallam	2	-	X	-	-
6	Dallam	5	-	X	X	X
7	Hangerő	1	X	X	X	X
8	Hangerő	3	-	-	X	-
9	Hangerő	5	-	X	X	X

semlegesítésre került. A harmadik fordulóban az elsőhöz hasonlóan mind a három jellemző szerephez jutott, de az egyéb beszédsajátosságokat rejtve hagyó, „vegytiszta” formában.

Az eredeti tervek között szerepelt még a jellemzők egyéb lehetséges kombinációinak (F0 és intenzitás, tempó és F0 stb.) prezentálása is, de ez jelentősen megnövelte volna a kísérleti alanyokra rótt terheket (ugyanazt a felvételt összesen tizenötször kellett volna végighallgatniuk), ami a rendelkezésre idő rövideje miatt sem fért bele a kísérlet kereteibe. A három feladatot három különböző alkalommal végezték el. Az egyes alkalmak időtartama így még az optimális 20 percen belül maradván (a 4 felvétel hosszából következően: ≈ 15 perc) biztosította folyamatos koncentrációjukat.

5.2.3. Az adatok feldolgozása

Az alanyokban kialakult hangérzetek alternatív leképezésére a reakciók bináris ábrázolása (növekszik, csökken) mellett egy másik, a nyers adatok konverzióján alapuló módszert is kidolgoztam, amely főként a géppel mért értékekkel való összevetést könnyítette meg. A eljárás lényege, hogy reakciók sorozata alapján egy olyan, a hangérzetben bekövetkező szintlépéseket kifejező görbét generál, amelynek formája a reakciók bináris értéke szerint változik a hangfelvétel időtengelyén. Pozitív reakció esetén egy értékkel nő, ellenkező esetben ugyanennyit csökken. A 5.4. ábra valós adatokon (feladat: dallam követése, kondíció: első felvétel első variációja) szemlélteti az eljárás kimenetét, ami tehát egyfajta rekonstrukciója a beszéd dallamáról kialakult mentális reprezentációnak. Magát az algoritmust a Praat program szkript nyelvén implementáltam. A *make_curve* eljárás a reakciókat idejét és előjelét tartalmazó TextGrid objektum (a forráskódban a reactionsID változó tárolja azonosítóját) kiválasztása után lép működésbe:

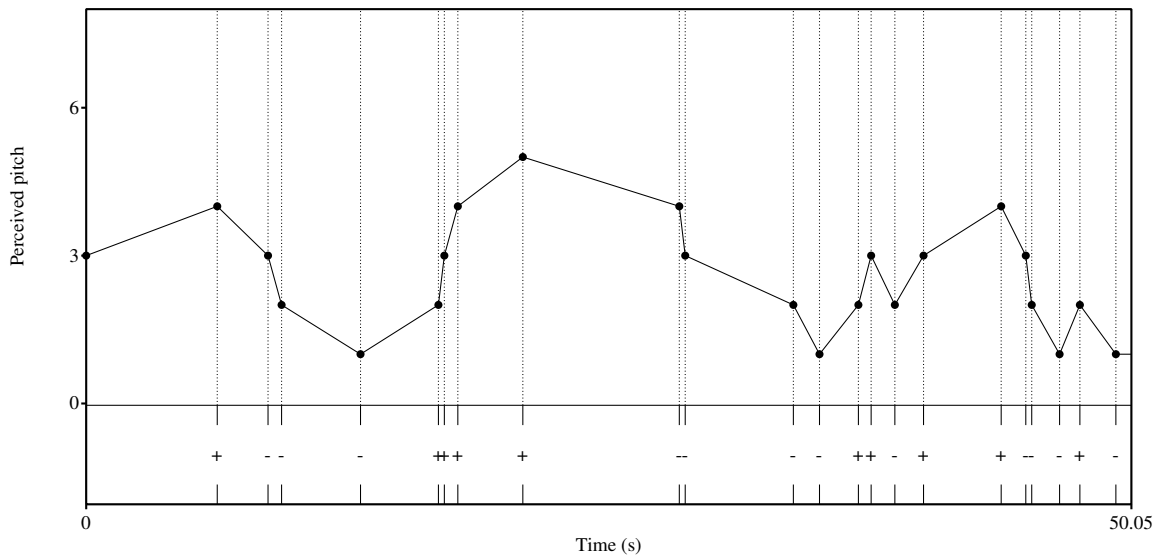

```

procedure make_curve
    baselevel = 3
    tdur = Get total duration
    points = Get number of points: 1
    curveID = Create PitchTier: name$, 0, tdur
    Add point: 0, baselevel
    selectObject: reactionsID
    level = baselevel
    for p from 1 to points
        time = Get time of point: 1, p
        direction$ = Get label of point: 1, p
        selectObject: curveID
        if direction$ = "+"
            level = level + 1
        else
            level = level - 1
        endif
        Add point: time, level
        selectObject: reactionsID
    endfor
endproc

```

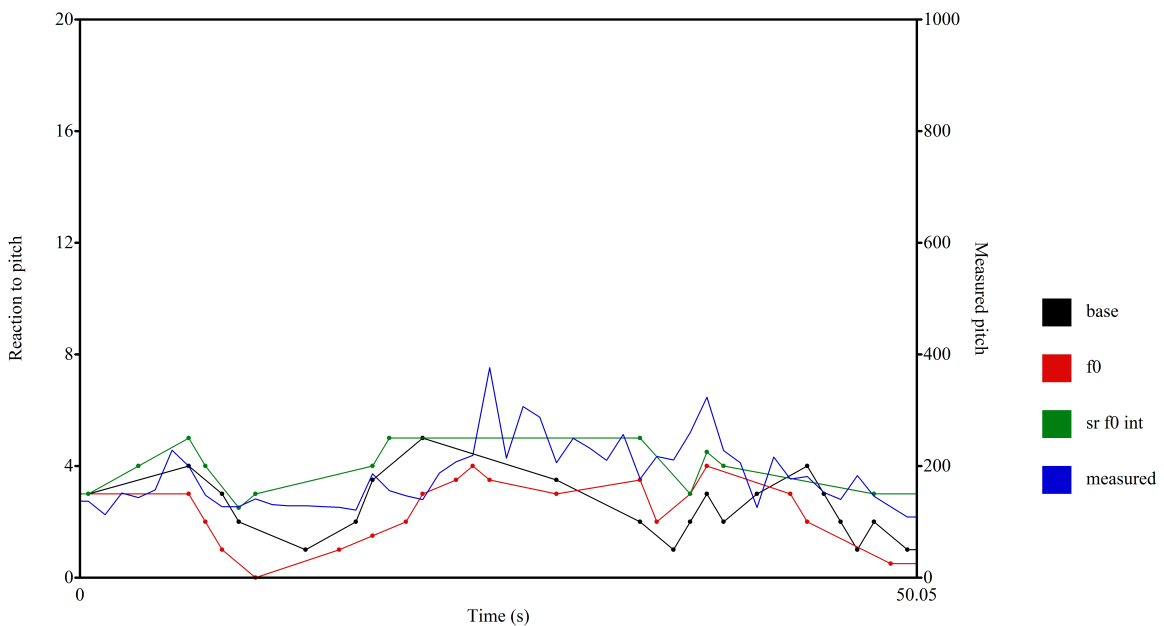
Mint a forráskódban látható, a kimenet egy „PitchTier” típusú objektumba kerül, ami a Praat programban eredetileg alapfrekvencia-adatok leképezésére szolgál, de kiválóan megfelel bármilyen idő-érték párokkal kifejezhető adatsor (intenzitás és tempó) tárolására is.

Az eredmények vizuális összevetésére háromféle – az ábrákat automatikusan generáló – rajzoló algoritmust dolgoztam ki. Az első bemenetűl azok a görbék szolgáltak, amelyek ugyanazon kísérleti alany eltérő kondíciók között (valamelyik dialógusrészlet különböző variációi) adott jelzéseit rögzítik (lásd 5.5. ábra). Ezek az ábrák a géppel mért adatokat (alapfrekvencia, intenzitás, szótagsűrűség) szemléltető görbékkel is kiegészültek, amelyeket előzetesen 500 milliszekundum hosszú ablakokra átlagoltam. A második típusba tartozó ábrák a különböző kísérleti alanyok azonos kondíciók között adott jelzéseinek összevetésére szolgáltak (5.6. ábra). A harmadik verzió bemenete az előző görbéinek összegzésével és 500 milliszekundum másod-



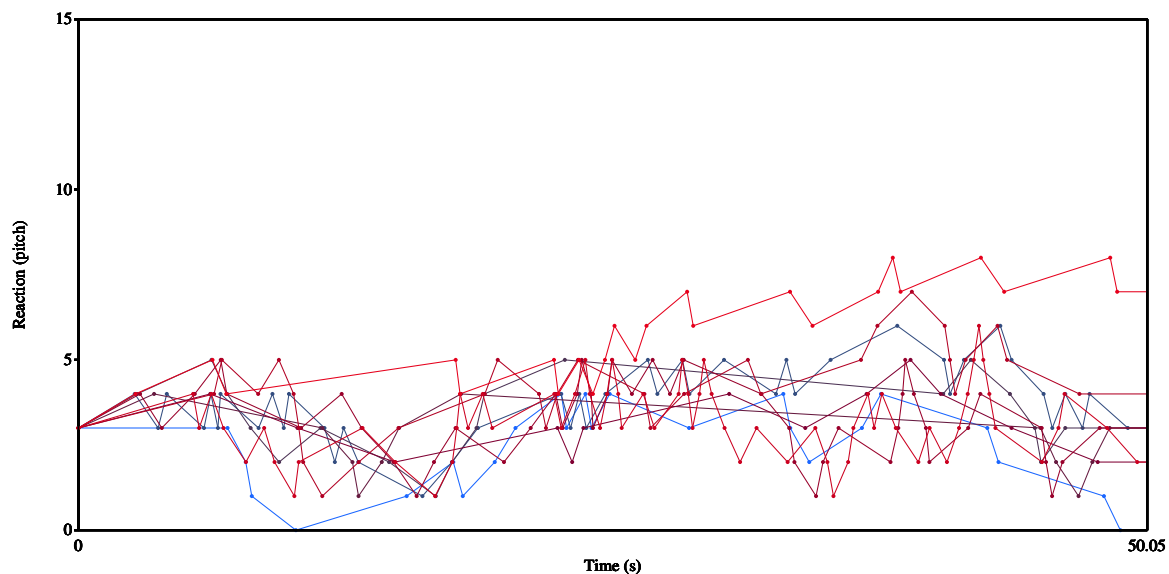
5.4. ábra. Az érzékelt intonáció szintértékeinek vizualizációja a kísérleti alanyok reakciói (a függőleges tengelyen látható „+” és „-” jelzések) alapján.

perc hosszú ablakokra átlagolásával az alanyok „átlagos” reakcióját hivatott vizualizálni a gépi mérésekkel való összehasonlításhoz (lásd 5.7. ábra).

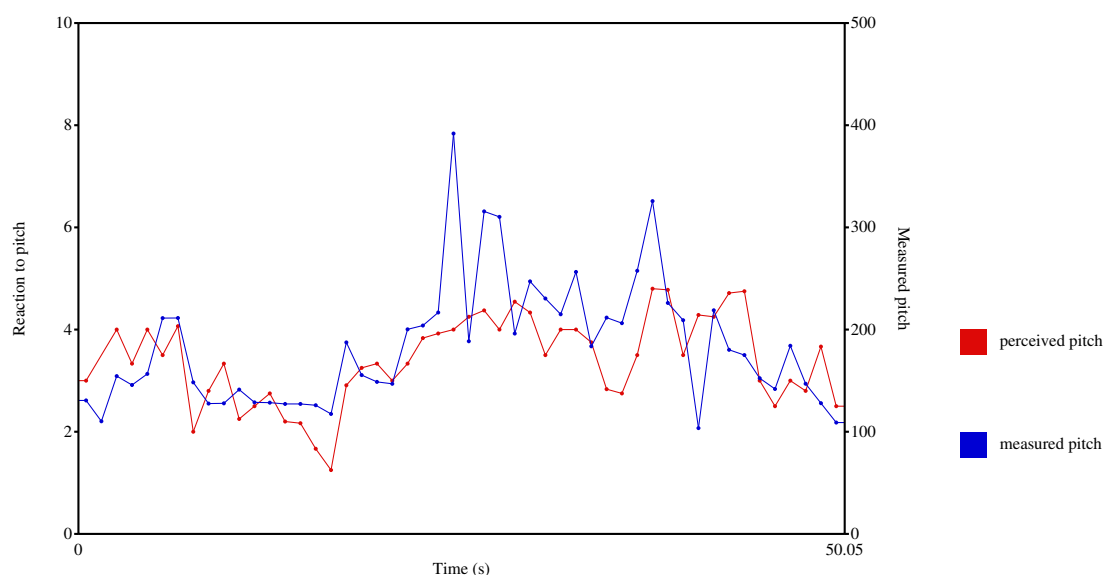


5.5. ábra. A hangmagasság változására különböző kondíciók között adott reakciók és a géppel mért alapfrekvencia adatok összevetése. A színjelzések feloldása felülről lefelé haladva: eredeti felvételre adott reakció (fekete), a felvétel 5. variációjára (minden jellemző együtt) adott reakciók (zöld), a felvétel 2. variációjára (csak F0, piros) adott reakciók, géppel mért alapfrekvencia adatok (kék).

Természetesen nem csak ábrák, de a különböző kondíciók között rögzített reakciók és a gépi mérések alapján olyan – az ábrákkal azonos rendezési elvet követő – adattáblák is készültek, amelyek a hangérzet ordinális skálán ábrázolt szintértékeit és a velük logaritmikus kapcsolatban



5.6. ábra. A hangmagasság változására azonos kondíciók (1. felvétel második variációja: csak az F0 hallható) között a különböző alanyoktól kapott válaszok.



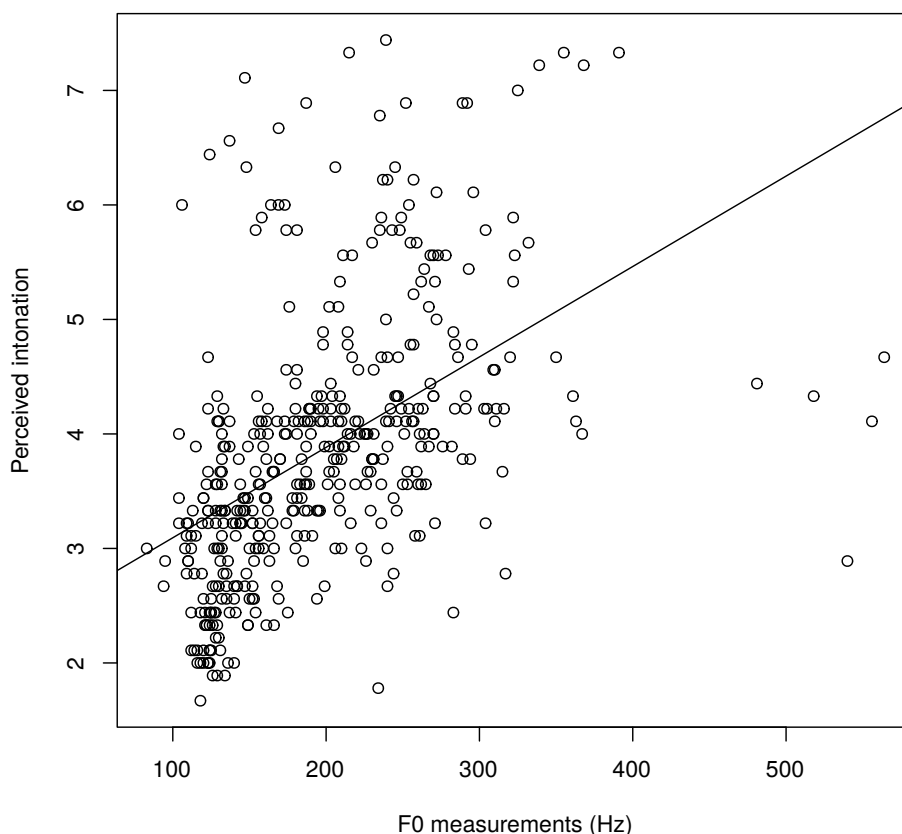
5.7. ábra. A hangmagasság változásaira adott reakciók 1 másodperc hosszú ablakokra átlagolt szintértékei az azonos módon feldolgozott alaphérfekvencia-adatokhoz viszonyítva.

álló⁹ mérési eredményeket 500 milliszekundum méretű ablakokra átlagolva állítják párhuzamba. Az adatsorokat később Spearman-féle rangkorrelációval vontam vizsgálat alá.

⁹ A Weber-Fechner-féle pszichofizikai törvény alapján az ingerként jelentkező, folytonos fizikai mennyiségek sokszorozódására az érzékelés ordinális skálákon való szintlépésekkel válaszol (Fechner, 1860).

5.2.4. Eredmények

Az alanyok szubjektív – a teszteteket követő rövid interjún¹⁰ adott – visszajelzését igazolva a beszéd szegmentális szerkezetétől megfosztott dallam (a felvételek Praat programmal újrászintetizált 2. variációja) követésénél adott reakciók alapján rekonstruált szintértékek korreláltak leginkább egymással és a géppel mért alapfrekvencia adatokkal. A szintértékek átlagát kifejező és a mérési eredményeket tartalmazó adatsorok (lásd 5.8. ábra) esetén a Spearman-féle korrelációs együttható értéke 0.61, ugyanez az eredeti felvétel követésénél jóval alacsonyabb ($r = 0.25$), a felvételek 5. verziójában pedig csak egyes alanyoknál volt bármiféle korreláció kimutatható. Az utóbbi esetben valószínűleg a tempót kifejező hangjelzések időbeli rövidsége nehezítette a hangmagasság megállapítását.

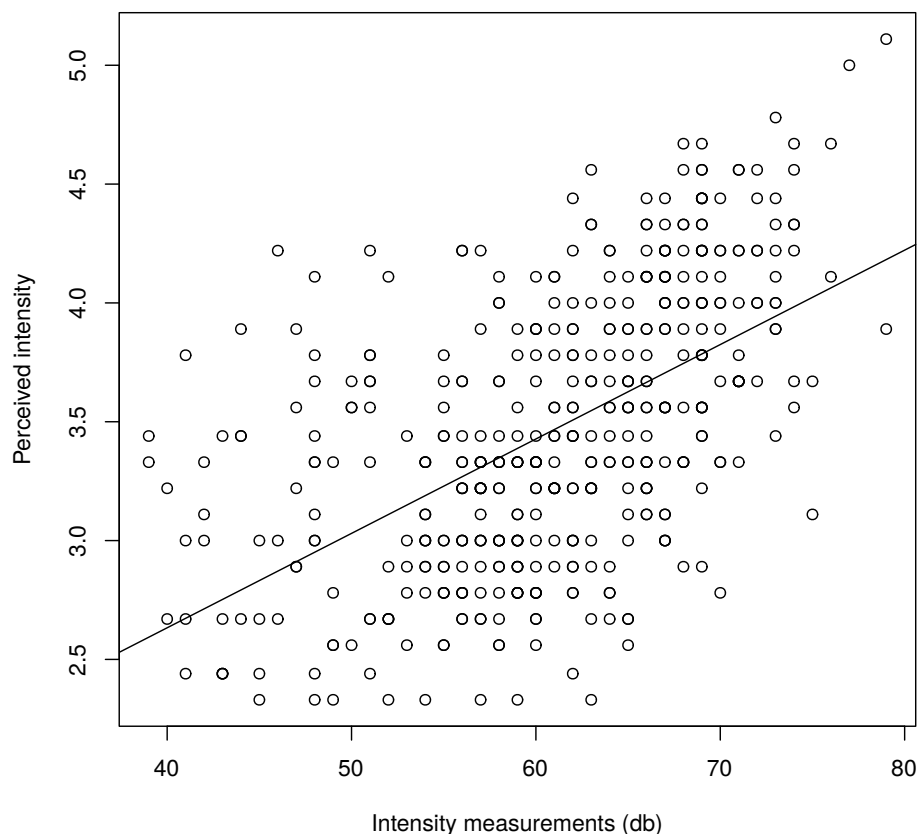


5.8. ábra. A mért és az érzékelt hangmagasság-változás közötti korreláció a felvételek 2. variációja esetén.

A beszédintenzitás változásainál érdekes módon a 3. fordulóban használt, a három jellemzőt együttesen prezentáló, az impulzusokat szótagonként szakaszoló felvételvariáció tűnt használ-

¹⁰ Informális keretek között arról kérdeztem meg őket, hogy a különböző feladatokat és kondíciókat összehasonlítva, melyiket érezték könnyebbnek vagy nehezebbnek.

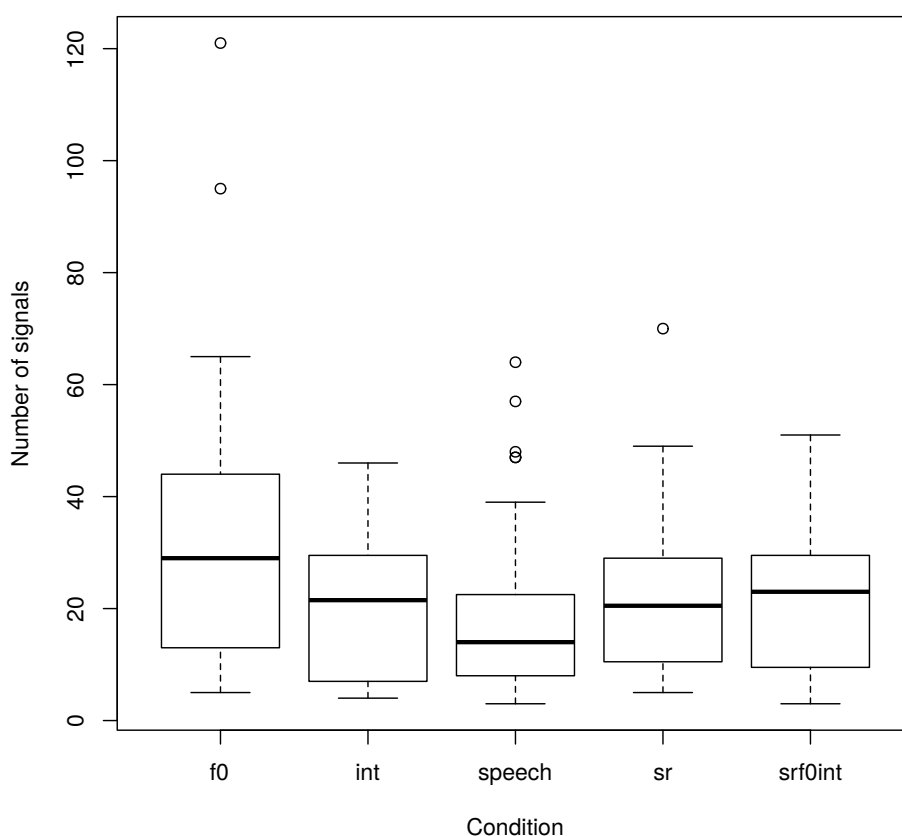
hatóbbnak az eredeti beszéd és a folyamatos, változó erősségű szinuszhullám helyett. A reakciókat vizsgálva legalábbis erre utal a gépi mérésekkel való magasabb, az 5.9. ábrán szemléltetett korreláció (a Spearman-féle együttható értéke: 0.58). Az eredeti felvételnél ugyanez 0.45, a szintetizált szinuszhullám esetén pedig elhanyagolható ($r = 0.25$), feltehetően azért, mert az intenzitás ingadozása az impulzus tagolatlansága miatt (izolálható szótagmagok hiányában) nem volt olyan kontrasztos, mint a másik két prezentációban.



5.9. ábra. A beszédintenzitás mért és az érzékelt változása közötti korreláció a felvételek 5. variációja esetén.

A beszédtempóra adott reakciók egymással korreláltak ugyan, de a szótagsűrűség alapján számolt adatokkal nem sikerült semmilyen kapcsolatot kimutatni. Ennek okát nem feltétlenül kell a feladat nehézségében keresni, sokkal valószínűbb, hogy a tempó változásainak mérésére kidolgozott módszerem (lásd 4.2.3. fejezet) a jellemző leképezésére nem volt megfelelő.

A kísérlet különböző kondíciók során rögzített eredményeit a percepcióban kialakult hangérzet részletessége (leadott jelzések száma) szerint jellemzi az 5.10. ábra. Ahol az alanyok gyakrabban éltek a jelzés lehetőségével, ott feltehetően jobban nyomon követhetőnek érezték a jellemzők különböző irányú változásait. A gépi mérésekkel mutatott erősebb korrelációkkal



5.10. ábra. Az alanyoktól kapott jelzések (szintlépések) átlagos mennyisége a különböző kondíciók esetén: eredeti felvétel („speech”), csak az alapfrekvencia (f0), csak az intenzitás (int), csak a tempó (sr), a három jellemző együtt (srf0int).

összhangban áll, hogy a legrészletesebb hangérzet a hangmagasság változásainak követésénél használt második felvételvariáció (az alapfrekvencia izolált prezentálására szolgáló stimulus) esetén volt rekonstruálható. Habár a jellemzők valós időben történő követése – azok gyakori változásai miatt – az összes kondíció esetén rendkívül nehéznek bizonyult, az alanyoktól kapott jelzések (szintlépések) átlagos mennyisége az eredeti beszédet rögzítő felvételek hallgatása közben volt a legalacsonyabb. Ez a korrelációktól függetlenül is érdekes tény általam nem ellenőrzött módon, de utalhat arra, hogy a kísérleti alanyokban az idegen nyelv használata ellenére is beindult egy magasabb szintű feldolgozási folyamat, amely megnehezítette a prozódiai jellemzők fizikai változásaira való koncentrációt. Ha tehát a felvételeken izolálható beszédhangokat és szavakat hallottak, akkor a percepcióban ösztönös módon a prozódiai jellemzők is a beszélők megnyilatkozásainak és attitűdjeinek értelmezését segítő nyelvi, illetve kommunikatív funkciója felől nyertek megvilágítást, amely a fizikai viszonyokról kialakult képzetek átrendeződését, perceptuális prioritásának csökkenését is maga után vonta.

5.3. „Angéla néni” esete

Az alábbi kísérletben néhány olyan, eredeti környezetéből izolált megnyilatkozás került a vizsgálat fókuszába, amelyek esetében feltételezhető volt, hogy az eltérő beszédhelyzetből következő attitűdtartalom a prozódiaiban a kifejezni kívántaknak leginkább megfelelő, Varga László kapcsolódó munkájában (Varga, 1994) *dallamprozodémának* nevezett sztereotip formákkal párosul. Mivel a kiválasztott megnyilatkozások egyéb tekintetben teljesen azonosak voltak egymással (az eredeti kontextusra semmilyen lexikális elem nem utalt vissza), egyedül szupraszegmentális sajátosságaik szolgálhattak megkülönböztető jegyként a különböző attitűdtartalmak beazonosításához, ami a megfelelő beszédhelyzet kiválasztásán keresztül a kísérleti alanyok feladata volt.

Az anyag összeállítását a karakterdallamok¹¹ Varga főntebb hivatkozott munkájában bemutatott háromszor hármass rendszere alapján végeztem el, ahol tehát mind a kilenc eset ugyanazon funkcionális mondat („Angéla néni”¹²) egy-egy megvalósulását jelentette különböző beszédhelyzetekben. A beszédhelyzetek közötti különbségeket rövid párbeszéd-variációkon keresztül világítottam meg, amelyekről egyetemi hallgatók közreműködésével hangfelvételek készültek. A szituációkat a kontextus előzetes tisztázása után párban játszották el, amelyek rögzítése az 5.1. fejezetben bemutatott felvételekkel azonos helyszínen és beállítások között történt. Habár a felvételeken hallható párbeszédek nem kezelhetők spontán beszédként (egy előre megírt szöveg felolvasása történt), az ominózus mondat („Angéla néni”) megformálására vonatkozóan nem adtam előzetes instrukciókat, mert a kiejtés természetességét helyeztem előtérbe az alább bemutatásra kerülő karakterdallamok pontos leképezése helyett, aminek sikerét megfelelő személyek hiányában eleve nem éreztem garantáltnak.

Az első három variáció az „eső jellegű” karakterdallamok jelentésének megvilágítását célozta. A csoport együttes jellemzésére Varga rendszerében az „önálló” címke szolgált. Ezen belül az „eső” karakterdallam a megnyilatkozások ‘önálló + befejezett’ címkét viselő típusát (lásd 1. példa) különíti el, ahol egy gondolatmenet a beszélő szándékai szerint lezárul, további kommentek hozzáadását az adott beszédszakaszon belül már nem igényli („Tegnap már megbeszéltük.”). Az ún. „féleső” karakter figyelhető meg azoknál az „önálló” megnyilatkozásoknál, ahol a gondolatmenet még egy további komment (vagy kommentek) hozzáadásával opcionáli-

¹¹ „Karakterdallam minden olyan, jelentéssel bíró hanglejtésforma, amelyben a hangmagasság vagy sztereotip módon változik, vagy magas szinttartó, és amelynek indító szótagja hangerőtöbblettel rendelkezik” (Varga, 1994)

¹² Varga által előszeretettel használt példa volt a karakterdallamok különböző jelentéseinek megvilágítására.

san, de folytatható („Ő mindig szívesen segít.”). Az „eső-emelkedő” karakter pedig azt jelzi, hogy a szóban forgó megnyilatkozás a beszélő érzése szerint valamilyen szempontból ellentétben áll („önálló + szembehelyezkedő”) a korábban vagy a később elhangzottakkal („Bár még nem kértem meg rá”).

- (1) – És ki fog a gyerekekre vigyázni amíg a moziban leszünk ?
– Tegnap már megbeszéltük. *Angéla néni*.
- (2) – És ki fog a gyerekekre vigyázni amíg a moziban leszünk ?
– *Angéla néni*. Ő mindig szívesen segít.
- (3) – És ki fog a gyerekekre vigyázni amíg a moziban leszünk ?
– *Angéla néni*. Bár még nem kértem meg rá.

A „lebegő jellegű” karakterdallamok jelentésének elkülönítésére szolgált a következő három variáció, amelyek közös jellemzője, hogy a kérdéses beszédszakasz „előre mutató” kérdésként igényli folytatását. Varga szavaival: „Olyanok ezek a dallamok, mint az előremutató ujj: nem egyszerűen azt jelzik, hogy a szóban forgó beszédszakasz folytatható, hanem egy közelgő önálló beszédszakaszra készítene fel, arra irányítják rá a figyelmet, annak a jelentőségét húzzák előre alá.” (Uo. 479. o.).

- (4) – Jó napot kívánok ! *Angéla néni* ?
– Angéla néni most nincs itthon. Délután felé szokott hazaérni.
- (5) – Jó napot ! *Angéla néni* ? Úgy volt, hogy ma itt találkozunk.
– Jó napot ! Angéla sajnos kórházba került. Én a bátyja vagyok.
- (6) – Szervusz Sanyi bácsi ! *Angéla néni* ?
– Szervusz, mindjárt szólok neki.

A három variáció közül a „magas szinttartó” karakterdallammal megvalósuló 4. tekinthető úgy-mond a semleges esetnek, amelytől az 5. variáció „feszült” kiegészítő címkével jellemezhető attitűdtartalmát az „emelkedő” karakterdallam fogja megkülönböztetni. Ez a „feszültség” a példaként szolgáló szituációban abból a meglepetésből ered, hogy az előzetes elvárásaimmal ellentétben, nem Angéla néni, hanem valaki más (számomra ismeretlen személy) jelenik meg az ajtóban. A 6. variációnál nem a feszültség, hanem éppen az „ereszkedő” karakterdallammal kifejezett „rutinszerűség”, ami a semleges esethez képest eltérést jelent, ha a hétköznapi megszokott állapotának megfelel, hogy Angéla néni otthon vár és Sanyi bácsi nyitja az ajtót.

Az „eldöntendő kérdő” csoportba tartozó utolsó három variáció esetében a „lebegő-eső” jelleg, ami fonológiai tekintetben érzékelhető különbséget jelöl. Közülük egyedül az ‘emelkedő-eső’ dallammal megvalósuló 7. variáció ruházható fel a szó szoros értelmében vett eldöntendő kérdésként (a kérdező ténylegesen választ vár) „kérdező” címkével. A másik két eset sokkal inkább a megütközés kifejezésére szolgál. A 8. variációban („szinttartó-eső”) még csak furcsálom („Reggel már beszéltünk”), a 9. esetben pedig már egyenesen lehetetlennek tartom, hogy a „kérdés” tárgyát képező proposíció igaz legyen.

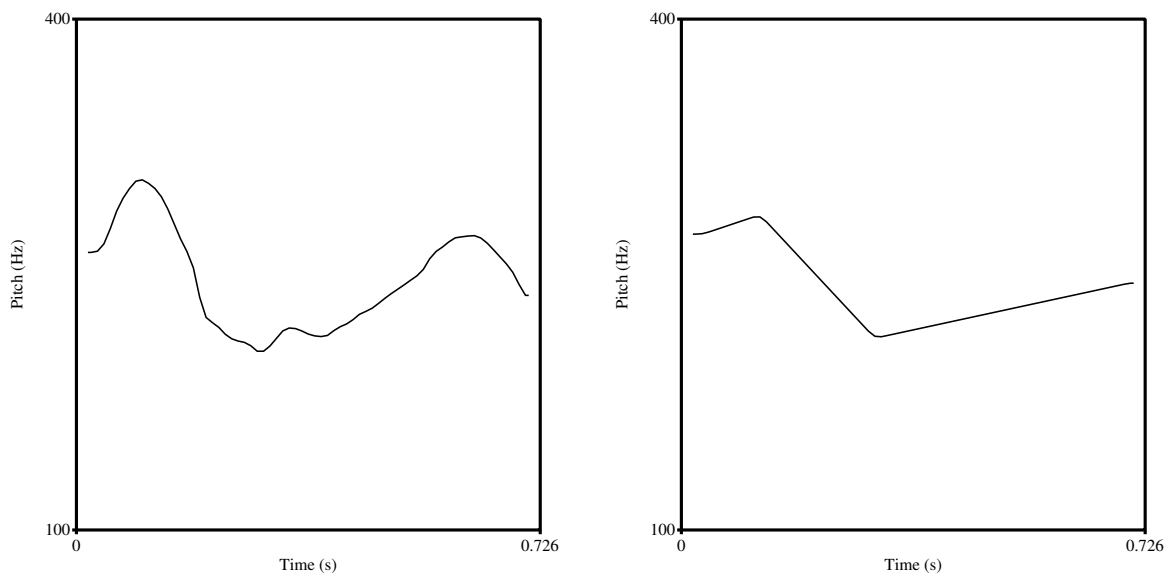
- (7) – Kerestek telefonon.
 – *Angéla néni?*
 – Igen, ő volt.
- (8) – Angéla néni keresett telefonon.
 – *Angéla néni?* Érdekes. Pedig reggel már beszéltünk.
- (9) – Angéla néni járt itt az előbb.
 – *Angéla néni?* De hisz ő most Japánban van.

A példák alapján készült hangfelvételekkel kapcsolatban fontos megjegyezni, hogy habár a kísérlet célja nem a beszédprodukció ellenőrzése (a modell dallamkontúrjainak a reprodukálása) volt, az alapfrekvencia mérések¹³ nem minden esetben igazolták a karakterdallamok főtebb felvázolt társulásait. Egyértelműen tetten érhetők voltak viszont a harmadik (önálló + szembehelyezkedő) és a hetedik (eldöntendő kérdő + kérdező) variációnál (lásd 5.11. és 5.12. ábra). Az elkészítés körülményei és a felvételek kis mennyisége miatt persze nyilvánvaló, hogy ezek a mérési eredmények nem képezhetik legitim alapját semmiféle, a karakterdallamok rendszerét érintő következtetésnek, aminek empirikus alapon való ellenőrzése a prozodémák és az általuk árnyalt, megnyilatkozásszintű jelentéstartalmak közötti speciális viszony miatt is problematikus. A karakterdallamok többértelműsége¹⁴, más tényezők vonatkozásában is megmutatkozó terheltsége (pl. a beszélő aktuális érzelmi állapota), illetve a megnyilatkozások előbeszédben sokszor tapasztalható, egyéb modalitások által kompenzált alul-formáltsága okán egyedül az intonáció szándékoltan vagy kórosan „hamis”, a kontextusnak *ellentmondó* eseteit tekinthetjük úgymond agrammatikusnak, nem odaillőnek. A percepció oldaláról persze továbbra is kérdés

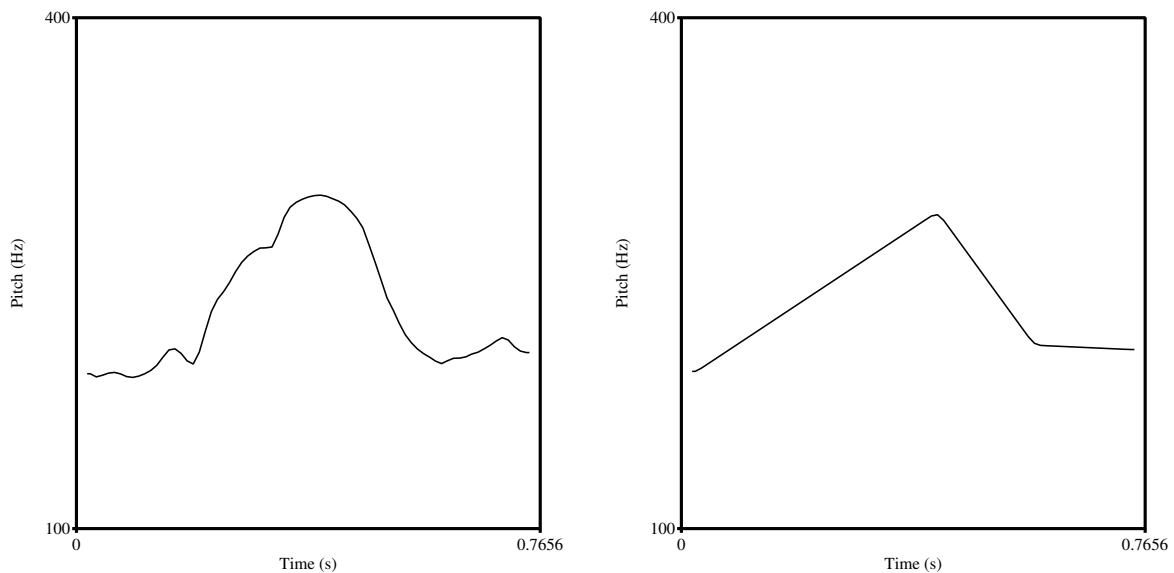
¹³ Az egyes variációkhoz tartozó F0-görbék ábráit a disszertáció függeléke rendszere tartalmazza.

¹⁴ Például az emelkedő-eső karakter nem csak eldöntendő, de ismételtető és az ún. visszhang kérdések esetében is érvényesül (Varga, 1994).

marad, hogy ezekben a stereotip modellhez képest alul-formált esetekben, a kontextus ismerete nélkül, mennyire lehetséges az attitűdtartalmak azonosítása, aminek ellenőrzése ugyebár a jelen kísérlet célja volt.



5.11. ábra. Az „Angéla néni” megnyilatkozás 3. variációja (önálló + szembehelyezkedő) alapján kapott F0-görbe (balra) és annak stilizált verziója (jobbra).



5.12. ábra. Az „Angéla néni” megnyilatkozás 7. variációja (eldöntendő kérdő + kérdező) alapján kapott F0 görbe (balra) és annak stilizált verziója (jobbra).

A kísérleti alanyok (9 nő és 3 férfi, átlagéletkor: 22 év) a fentebbi modellt nem ismerték, és komolyabb nyelvészeti háttérrel sem rendelkeztek. A teszt az előző kísérletekhez hasonlóan egy webes interfész¹⁵ használatával történt, aminek első fordulójában az eltérő beszédhelyzeteket

¹⁵ <http://lingua.arts.unideb.hu/angela/index.html>

reprezentáló dialógusokból izolált megnyilatkozásokat kellett véletlenszerű sorrendben fejhallgatóval meghallgatniuk és a legmegfelelőbbnek tűnő szöveges kontextusba visszahelyezniük. A felvételeken végig ugyanaz a beszélő hallható, a meghallgatások számát nem korlátoztam, a kilenc választási lehetőség pedig minden döntés esetén rendelkezésre állt, azaz előfordulhatott, hogy ugyanahhoz a kontextushoz több felvétel is hozzárendelésre került.

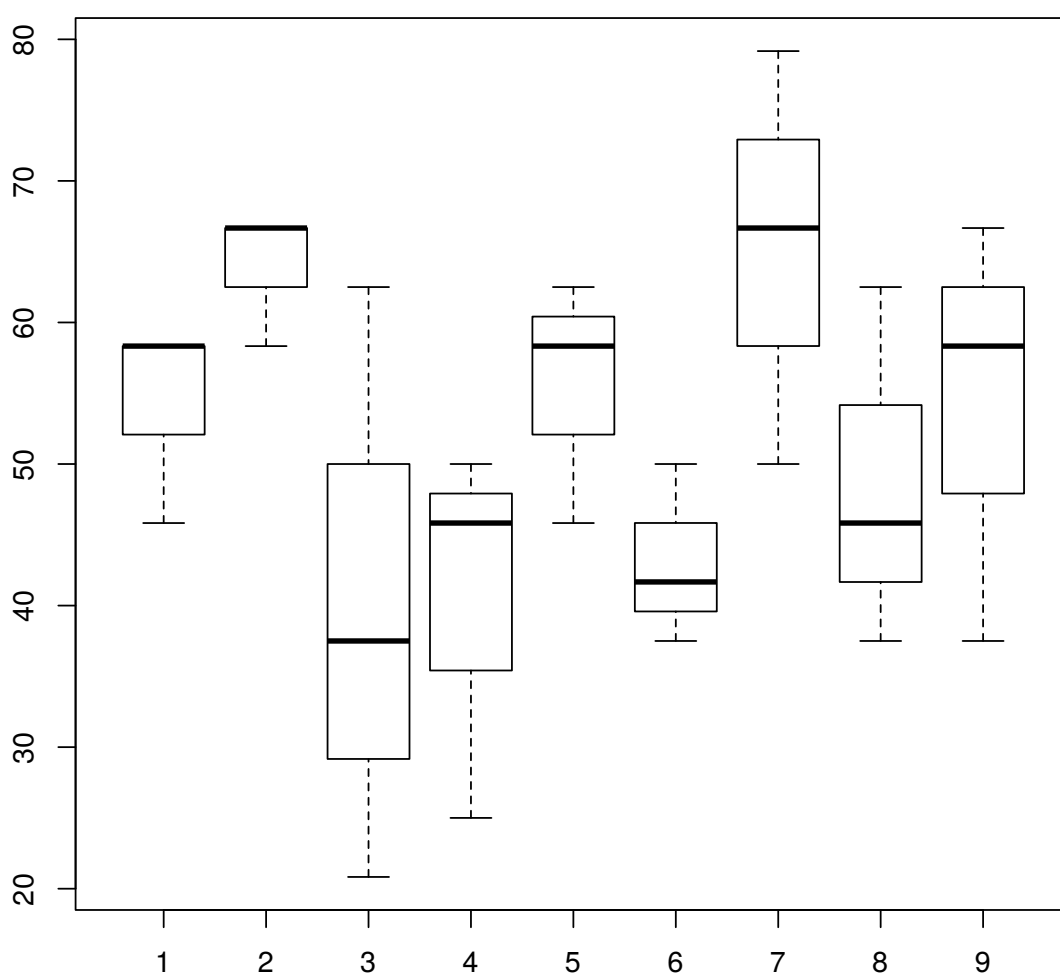
Mivel az egyes döntéseket a hanglejtésen kívül a megnyilatkozások egyéb prozódiai jellemzői (pl. az eltérő hangszínezet) is befolyásolhatták, indokoltnak láttam egy második forduló beiktatását, amelyben ezeket az egyéb tényezőket kizárva a beszélő hangját a felvételekből mért alapfrekvencia-értékek alapján generált, a megnyilatkozás eredeti hanglejtését pontosan lekövető szinuszhullámmal¹⁶ helyettesítettem. Ezután következett még egy harmadik forduló, amiben az azonos módon előállított hullámok nem az eredeti, hanem az alapfrekvencia-értékek előzetes simítása és stilizálása után kapott görbék hangmagasság-változásait reprezentálták. A felvételek utóbbi módon átdolgozott verziója (lásd. az 5.11. és az 5.12. ábra jobb oldalát) a hanglejtés stilizálásával tulajdonképpen a karakterdallamok sztereotip, fonológiai releváns formáira próbálta az eredeti görbéket (mint az előbbiek fonetikai megvalósulásait) úgymond visszaredukálni. A kísérletben való használatukkal azt próbáltam ellenőrizni, hogy a stilizálás eredményezte információvesztés az alanyok számára nehezíti-e a megnyilatkozás attitűdtartalmának beazonosítását. Más szóval: az átalakítással kiiktatott fonetikai részletek mennyire voltak relevánsak vagy irrelevánsak a percepció számára. Mivel a 4.2. fejezetben bemutatott ProsoTool algoritmusában ugyanezeket az eljárásokat használtam fel, a kísérlet kimenetele a disszertáció keretében megvalósított fejlesztések ismételt ellenőrzése szempontjából is jelentőséggel bír.

A különböző attitűdtartalmak háromszor hármas rendszeréből következik, hogy az azonos csoportba („önálló”, „előre-mutató” és „eldöntendő-kérdő”) tartozó megnyilatkozások nehezebben elkülöníthetők, mivel az attitűdtartalmuk csak egy-egy árnyalatban különbözik egymástól. Az egyes döntések értékelésére ezért egy három szintű skálát vezettem be, ahol 2 ponttal értékeltem a teljesen pontos azonosítást, 1 ponttal pedig azokat az eseteket, amelyek a megnyilatkozást a megfelelő csoportba helyezték vissza (csak az „árnyalatot” illetően történt tévedés), ellenkező esetben a válasz 0 pontot ért. Az eredmények összegzése kétféle csoportosítási szempontot követett: százalékos alapon vizsgáltam az egyes alanyok teljesítményét (hány pontot értek el a maximum 18-ból), illetve az egyes variációkra vonatkozó, sikeres azonosítások arányát (hány pont gyűlt össze az értékelésekből) mind a három kondícióban (beszéd, részletes és stilizált F0).

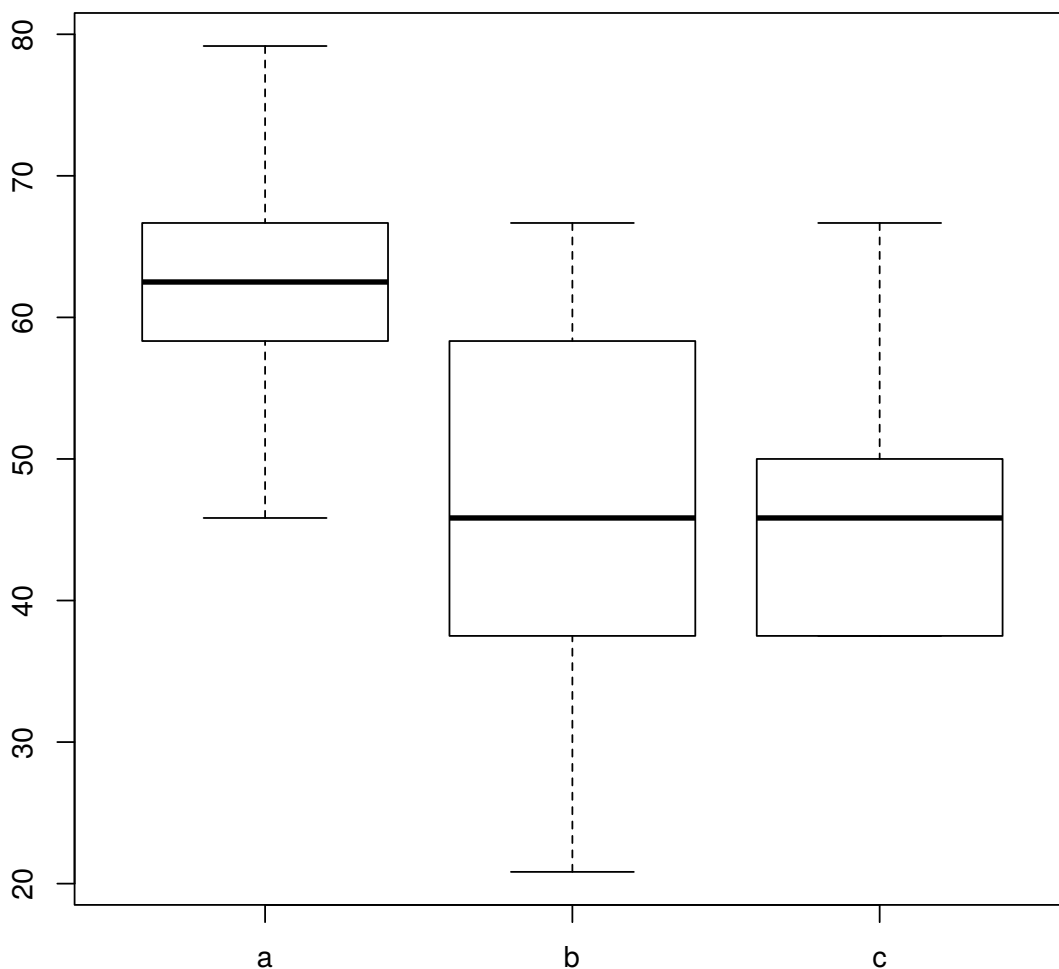
¹⁶ Ezek a hangfájlok az előző fejezetben bemutatott kísérlethez hasonlóan a Praat program eljárásaival („To Sound (sine)...”) készültek.

5.9. táblázat. Az attitűdtartalmak azonosítási rátája a különböző lejátszási módokban.

Variáció	Eredeti felvétel	részletes F0	Stilizált F0
1	58%	58%	45%
2	66%	58%	66%
3	62%	20%	37%
4	45%	25%	50%
5	58%	62%	45%
6	50%	41%	37%
7	79%	66%	50%
8	62%	45%	37%
9	66%	37%	58%

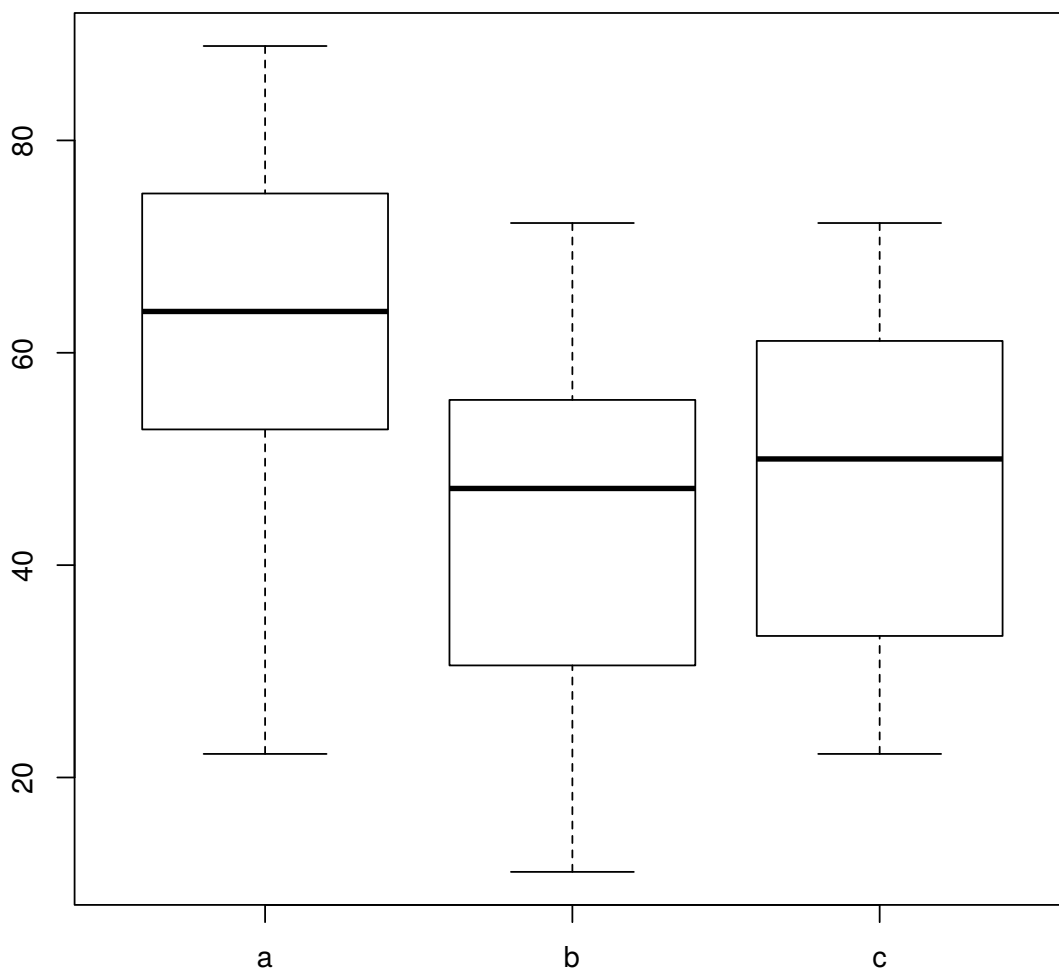


5.13. ábra. A kontextus beazonosításának sikeressége a megnyilatkozások különböző variációinak (1-9) függvényében. Mivel minden variáció 3 verzióban került tesztelésre, a szórások a különböző verziókra számolt találatok ingadozását mutatják.



5.14. ábra. A kontextus beazonosításának sikeressége a felvétel variációjának (a-c) függvényében: (a) eredeti felvétel, (b) részletes F0, (c) stilizált F0.

Mint a 5.9. táblázatban és a 5.13 ábrán látható, a kilenc variáció közül az azonosíthatóság tekintetében az emelkedő-eső dallammal megvalósított 7. variáció („eldöntendő-kérdő + kérdező”) az, ami leginkább kiemelkedik, habár a különböző felvételverziókra nézve ez jelentősebb szórást mutat. Az „eső jellegű” csoportba tartozó 1. és 2. variációk esetében – a 5.16. és 5.17. ábrán látható jóval laposabb megvalósulásuk ellenére – a szintetizált verziók alapján kapott eredmények kevésbé maradnak el az eredeti felvétel mögött. Erre egy lehetséges magyarázattal szolgálhat a megnyilatkozások karakterdallamának („eső” és „féleső”) élesebb elkülönülése: a többivel ellentétben sem irányváltás, sem pedig emelkedés nem tapasztalható bennük. Érdekes, hogy habár az egymással való hibás felcserélésük igen gyakori volt (összesen 11 alkalom), a két eset sikeres elkülönítésére még a stilizált alapfrekvencia esetén is bőven akad példa (a 2. va-

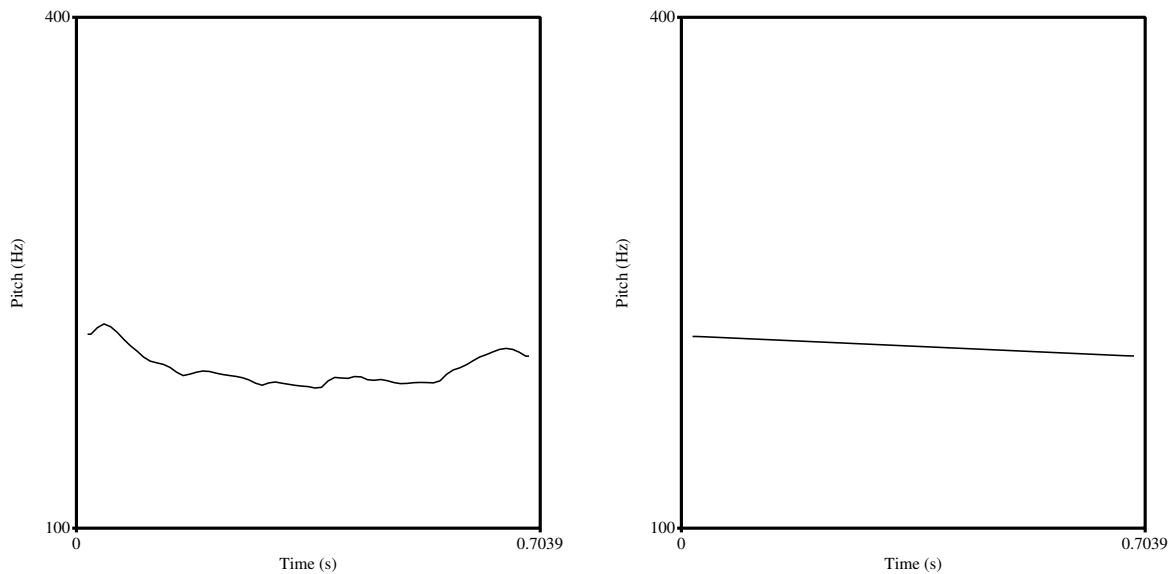


5.15. ábra. A kísérleti alanyok teljesítménye a felvétel variációjának (a-c) függvényében: (a) eredeti felvétel, (b) részletes F0, (c) stilizált F0.

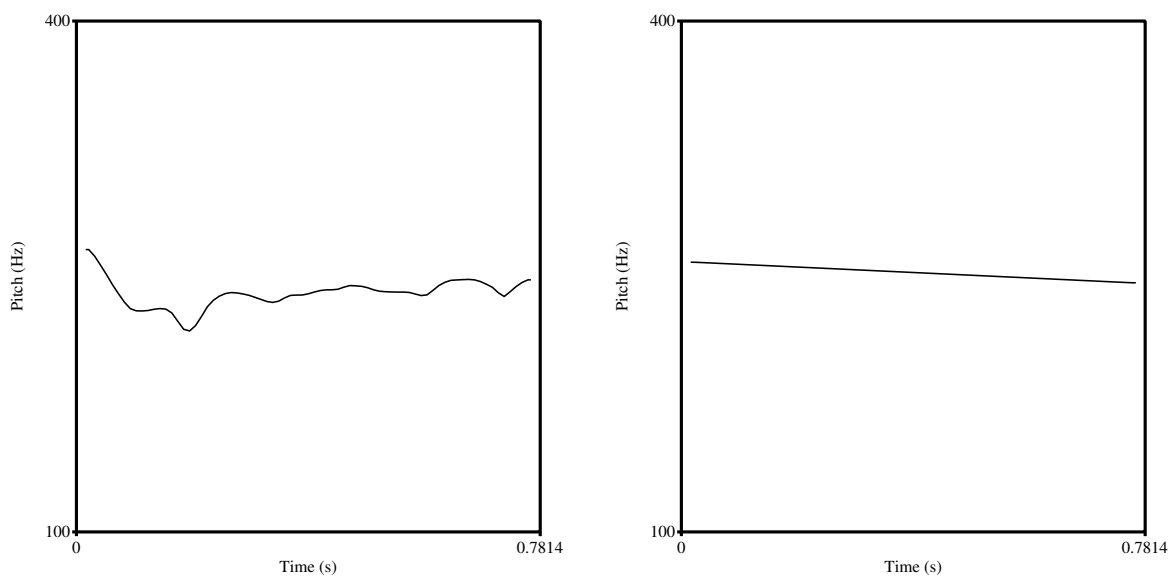
riációt a próbálkozók 58%-a sikeresen azonosította), amit Varga László nyomán a két karakter záróhangmagasságának relatív (az ábrákon is megfigyelhető) különbsége magyarázhat:

„A hallgató általában rövid idő alatt «beméri» a beszélő alapvonalát, és meg tudja állapítani, hogy a beszélő elérte-e azt vagy sem. Ezért az eső és félig eső karakterek különbségét kategorikus különbségnek tartjuk.” (Varga, 1994. 477. o.)

A tévesztések leggyakoribb esete a 4. és a 6. variáció felcserélése volt (összesen 15 alkalom), ami nagy valószínűséggel a két megnyilatkozás megvalósításának, illetve a párbeszédes szituációk hasonlóságának számlájára írandó, míg az azonos csoportba tartozó 5. felvétel elkülönülését a feszültséget jobban kidomborító szituáció segíthette. Ugyanez elmondható a harmadik csoportban szélső értéket képviselő 9. megnyilatkozásáról is. A legkevésbé sem érthető



5.16. ábra. Az „Angéla néni” megnyilatkozás 1. variációja (önálló + befejezett) alapján kapott F0-görbe (balra) és annak stilizált verziója (jobbra).



5.17. ábra. Az „Angéla néni” megnyilatkozás 2. variációja (önálló) alapján kapott F0-görbe (balra) és annak stilizált verziója (jobbra)

viszont, hogy a 3. variációnál („önálló + szembehelyezkedő”) a teljesen prototipikus módon megvalósított hanglejtés ellenére miért maradtak ennyire alul a szintetizált verziók az eredeti-hez képest.

A „Mit mennyien?” és „Ki mennyit?” kérdéseket az eltérő felvételverziók szerint összegezõ 5.14. és 5.15. ábra alapján megállapítható, hogy megfelelő kontextus társítása az eredeti felvételek hallgatása közben sikeresebb volt. Ez igazolni látszik azt a kezdeti feltevést, miszerint az attitűdtartalmak beazonosítását a hanglejtésen kívül a megnyilatkozások egyéb prozódiai jellemzői is pozitívan befolyásolhatják. Itt elsõsorban az eltérõ hangszínézetre gyanakodhatunk, de

az egyik kísérleti alany szubjektív visszajelzésében hivatkozott a „bizonytalanabb intonációra” és a „tempó felpörgetésére” is. Az intonáció automatikus annotációjához alkalmazott eljárások validációja szempontjából kedvező eredmény, hogy az alaphfrekvencia prezentálására használt két szintetizált verzió (stilizált és nem stilizált) között nem volt hasonló különbség tapasztalható sem az egyes variációk azonosíthatóságának, sem pedig az egyes alanyok teljesítményének tekintetében. A három kondícióban (eredeti felvétel, stilizált és nem stilizált F0) kapott eredmények közötti különbséget mind a 9 hanglejtés-variációra Friedman próbával ellenőriztem, ami csak a harmadik (önálló + szembehelyezkedő), illetve a kilencedik (kérdő + felkiáltó) esetben mutatott ki szignifikáns különbséget az eredeti felvétel és a stilizálás nélkül szintetizált F0 alapján kapott eredmények között. A Friedman próba eredménye a 3. variáció esetén:

$$\chi^2(2, N = 3) = 6.3429, p = 0.04194 \quad (5.1)$$

A Wilcoxon próba eredménye ugyanitt az a) és a b) kondíció között:

$$V = 28, p = 0.01519 \quad (5.2)$$

A Friedman próba eredménye a 9. variáció esetén:

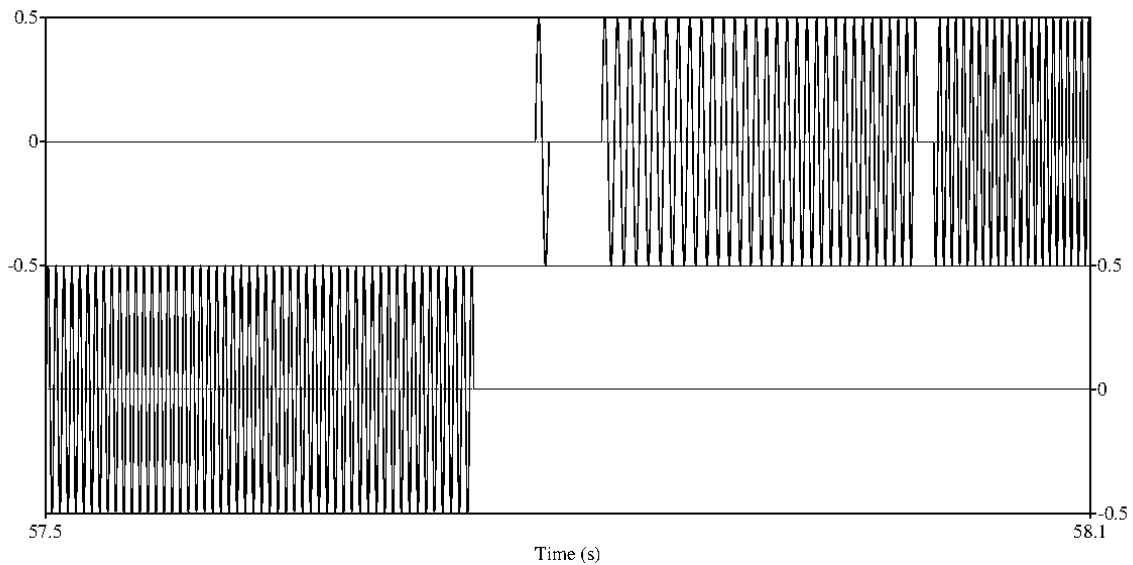
$$\chi^2(2, N = 3) = 7.4, p = 0.02472 \quad (5.3)$$

A Wilcoxon próba eredmény az a) és a b) kondíció között:

$$V = 28, p = 0.01141 \quad (5.4)$$

5.4. Formális vagy informális?

A utolsóként bemutatandó percepciók kísérlet az 5.1. fejezetben ismertetett teszt első kérdéséhez („irányított” vagy „kötetlen”) hasonló feladat elé állította a résztvevőket, akiknek hangfelvételek alapján kellett dialógusrészleteket kategorizálniuk abból a szempontból, hogy az illető részlet egy formális vagy pedig egy informális társalgáshoz tartozik-e. Az korábbtól eltérő kondíciók használatát egy a 6.1. fejezetben bemutatásra kerülő, azonos céllal, de mély neuronhálókkal végzett másik kísérlet motiválta. Össze akartam ugyanis hasonlítani a közel azonos „bemenet” alapján hozott gépi és emberi reakciók eredményeit.



5.18. ábra. Az adatközlő (fent) és az interjúvezető (lent) hangját helyettesítő, változó frekvenciájú szinuszhullámok az eredeti felvételek újrasszintetizált verzióiban.

A kísérlet anyagaként a HuComTech korpusz szimulált állásinterjúit és az ezeket követő (azonos adatközlővel zajló) informális beszélgetéseket rögzítő hangfelvételek újrasszintetizált, csak a fordulóváltás és az intonáció szerkezetét reprezentáló variációit használtam fel. Az újrasszintetizálás célja az volt, hogy a kísérleti alanyok ne a társalgások verbális tartalma (az interjúban ismert nyelven elhangzó szavak alapján túlságosan egyszerű lett volna a feladat), hanem a mély neuronhálókkal végzett klasszifikációhoz is használt jellemzők alapján hozzanak döntéseket. A technikai kivitelezés szintjén ez olyan (az 5.2. fejezetben leírtakhoz hasonló) hangfájlok generálását jelentette, amelyek a beszélőket külön csatornán elszeparálva, az eredeti F0-görbék frekvenciáját követő szinuszhullámok segítségével fejezték ki a fordulóváltások helyét, illetve a hanglejtés változásait. Az 5.18. ábrán látható részlet épp egy fordulóváltás pillanatát szemlélteti, ahol a 2. csatornán elhelyezkedő, az interjúvezető hangját helyettesítő szinuszhullámra az 1. csatornán egy alacsonyabb, de növekvő frekvenciájú, az adatközlő reakcióját kifejező hang válaszol.

A kísérlet 9 egyetemi hallgató részvételével bonyolódott le, akik összesen 216 darab (108 formális, 108 informális), később a mély neuronháló tesztalmazásában szereplő dialógusrészlet értékelését végezték el. Az anyagot három egyenlő részre osztottam fel közöttük, ami személyenként 72, felvételenként pedig összesen három értékelést jelentett. A túlságosan hosszú koncentrációt kerülendő a stimulusok három fordulóban (3 x 24 felvétel) kerültek lejátszásra. Egy felvételnél átlagosan 30 másodperc belehallgatás volt szükséges ahhoz, hogy az alanyok döntésre jussanak. A teljes anyagra kapott többségi döntések eredménye pedig 56 százalékban állt

fedésben a dialógusrészletek tényleges hovatarozásával. Ennek a vártnál negatívabb eredménynek az értelmezésére később, a mély neuronhálókkal végzett kísérlet ismertetése után fogunk visszatérni.

6. FEJEZET

Kísérletek gépi tanulással

Az alábbi fejezet a disszertáció azon kísérleteit foglalja össze, amelyekben a spontán beszéd prozódiai jellemzőit már nem az emberi percepció, hanem a gépi tanulás bevonásával vizsgáljuk tovább, és a mesterséges intelligenciát állítjuk az előző fejezetben bemutatott tesztekhez hasonló feladatok elé. A cél egyik kísérlet esetében sem az összes elérhető jellemzőt felhasználó és a maximális hatékonyságot garantáló eljárás kivitelezése volt, hanem annak a vizsgálata, hogy a szemantika és a vizualitás kizárásával, pusztán prozódia információkra támaszkodva milyen eredmények és „döntések” születnek egy, a tulajdonképpeni kérdésre és tartalomra „vak” rendszeren belül. Ugyanakkor a vizuális és szöveges információk mellőzését – a hipotézisvizsgálat teoretikus megfontolásain túl – a gyakorlati felhasználás vonatkozásában is indokolhatjuk azokkal a korántsem rendkívüli szituációkkal, melyekben – az adott nyelvre implementált beszéd-felismerő rendszer, illetve a vizuális csatorna hiányában – nincs lehetőség egyéb modalitások bevonására. A kísérletek bemutatása előtt fontosnak érzem kiemelni, hogy megvalósításukban nélkülözhetetlen segítséget jelentett Kovács Görgy kollégám munkája, aki az osztályzási feladatokra felhasznált neuronhálók kialakítását és betanítását saját munkakörnyezetében (Abadi et al., 2015) végezte el. A következő alfejezetben bemutatásra kerülő kísérlet pontos technikai részletei egy korábban megjelent konferenciaközleményünkben (Szekrényes & Kovács, 2017), a témaváltások automatikus detektálásáról szóló munkánkat pedig könyvfejezetként (Kovács & Szekrényes, 2019) publikáltuk.

6.1. Formális és informális dialógusok megkülönböztetése

Az alábbi kísérlet kettős céllal került kidolgozásra. Egyrészt a disszertáció egyik hipotézisének az ellenőrzésére, másrészt a prozódia automatikus annotációjára szolgáló eszköz hatékonyságának, a kimenetként kapott címkék gépi tanulásban való alkalmazhatóságának a tesztelésére.

Az ellenőrizni kívánt hipotézis szerint egy társas interakció formális időviszonyai és prozódiai sajátosságai alapján a társalgási szituáció – mint mögöttes információ – is hatékonyan rekonstruálható, ha egy tárgyalási univerzumon belül kellően le tudjuk szűkíteni annak lehetséges, formálisan megragadható kategóriáit. A kísérletben kitűzött osztályzási feladat egészen konkrétan az 5.4. fejezetben bemutatott percepció teszt mély neuronhálókkal való reprodukciója, amelyben ezúttal egy tanuló algoritmusnak kellett a beszélők intonációja és a beszélőváltás ritmusa alapján eldöntenie, hogy egy interakció formális vagy informális dialógusok halmazába tartozik-e.

6.1.1. Az osztályzás anyaga

Az osztályzási feladat anyagaként a HuComTech korpusz szimulált állásinterjút (formális részhalmaz: 111 darab) és az interjút követő informális beszélgetéseket (informális részhalmaz: 111 darab) használtuk fel. Mint az 5.4. fejezetben már kitértünk erre, a két csoport közötti különbségtételt erősen megnehezíti, hogy a dialógusok mind a két esetben az interjúvezető irányítása alatt, egy előre megtervezett kérdéssor használatával zajlottak le. Fontos különbség viszont, hogy az állásinterjúk egy jóval merevebb, távolságtartó beszédmódban, az informális beszélgetésekhez képest feszültebb hangulatban zajlottak. Az utóbbiak oldottságához, közvetlenebb hangneméhez az interjú során szóba kerülő témák személyes jellege mellett a 4.1. fejezetben tárgyalt pszichológiai tényezők is hozzájárultak.

Hogy megnöveljük a tanításhoz szükséges példák számát, az egyes interjút további modulokra, az interjúkérdések mentén dialógus-témákra daraboltuk fel. Mivel az állásinterjúkban több kérdés hangzott el, mint az informális dialógusokban, a modulok további szűrése révén igyekeztünk kiegyenlíteni a részhalmazokra eső számbeli különbséget. A szűrés során kihagyásra kerültek azok az adott téma szempontjából kifejtetlen dialógusrészletek, amelyek időbeli rövidsége ($t < 30\text{sec}$) alkalmatlanná tette őket arra, hogy felhasználhatók legyenek a neuronhálók tanítása során. Szintén kihagyásra kerültek az interjúk bevezető, a tényleges témákon kívül eső szakaszai. A szűrések eredményeként 799 formális és 808 informális példa állt rendelkezésünkre. Mint a 6.1 táblázatban látható, a formális és az informális dialógusrészletek időbeli hossza jelentős eltérést mutat. A tanítás során ezért fontos szempontot jelentett, hogy erre az esetleges tényezőre a neuronhálók ne támaszkodjanak és csak a lokális időviszonyokat (beszédfordulók és dallamszegmentumok időtartama) kapják bemenetként.

6.1. táblázat. Az interjútemák darabszáma, teljes és átlagos időtartama a formális és informális dialógusokban (előzetes szűrés után).

	Darabszám	Teljes időtartam (s)	Átlagos időtartam (s)
Formális	797	52549	65.93
Informális	713	101103	141.8

6.1.2. Felhasznált jellemzők

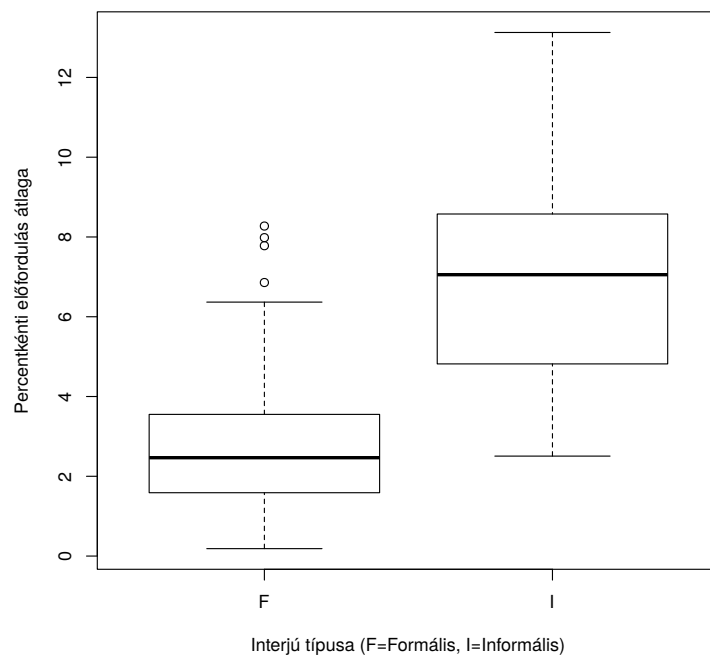
6.1.2.1. Beszélőváltás

A beszélőváltás annotációjának a kísérletben történő felhasználását két megfontolás motiválta. Egyrészt a ProsoTool algoritmus a beszélők egyedi hangterjedelmének meghatározásához és a dallamkontúrok címkézéséhez eleve bemeneti követelményként támaszkodott ezekre, a kimenetben is explicite szereplő információkra. Másrészt feltételeztük, hogy a beszélőváltás szekvenciái, a beszédfordulók ritmusa – mint tágabb értelemben vett prozódiai jellemző – az osztályzási feladatban önmagában is fontos, a korpusz formális és informális felvételeinek tekintetében megkülönböztető tényezőt jelenthet. Mint a 6.2. táblázatban és a 6.1 ábrán látható, az előzetes számítások alapján az egyes beszédfordulók és szünetek átlagos időtartama és gyakorisága (percenkénti előfordulás) nem, egyedül az átfedő beszéd gyakorisága mutatott szignifikáns különbséget (a páros t-próba eredménye: $p = 2.2e - 16$) a formális és az informális felvételek között. Az informális felvételek esetében közel háromszoros arányban fordult elő. Ez a megfigyelés visszaigazolja azt az intuíciót, miszerint az állásinterjú kötöttebb formájához képest az informális dialógusokban a beszélők viselkedése kevésbé volt feszélyezett, gyakrabban fordultak elő közbeszólások, a figyelem folytonosságát megerősítő visszajelzések, esetenként átfedő beszéddel kezdődő fordulóátvételek.

6.2. táblázat. A különböző kategóriájú beszédsegmensek (interjúalany, interjúvezető, átfedő beszéd) és szünetek átlagos előfordulása (forduló / perc) és időtartama (másodpercben) a formális és az informális dialógusokban.

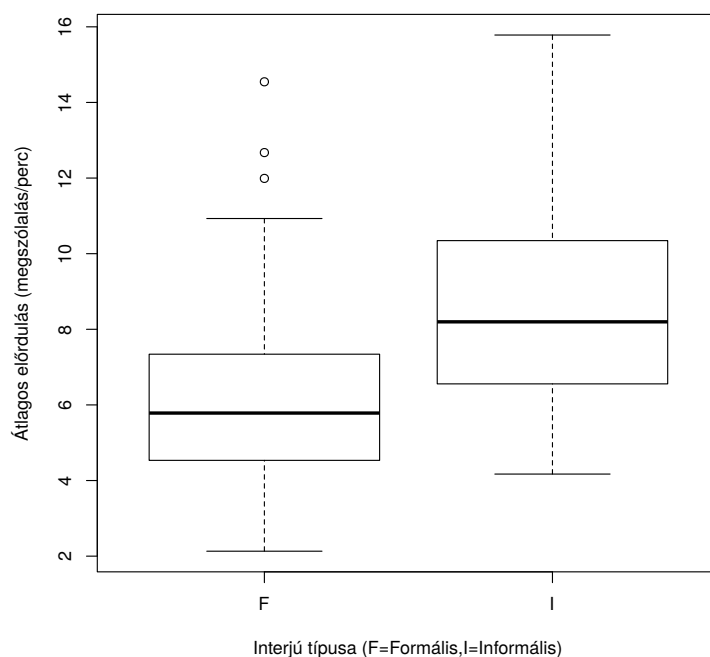
	Részhalmaz	Adatközlő	Interjúvezető	Átfedő	Szünet
Előfordulás	Formális	9.25	5.59	2.48	10.43
	Informális	9.88	8.52	6.32	8.99
Időtartam	Formális	4.77 s	3.02 s	0.53 s	0.77 s
	Informális	3.09 s	2.55 s	0.76 s	0.67 s

Az adatok alapján megfigyelhető másik, kevésbé jelentős különbség, hogy az interjúvezető megszólalásainak gyakorisága (6.2) az informális dialógusokban fokozott aktivitásra utaló növekedést mutat, ami az interakciók személyesebb hangvételével, az alá- és fölé-rendeltségi



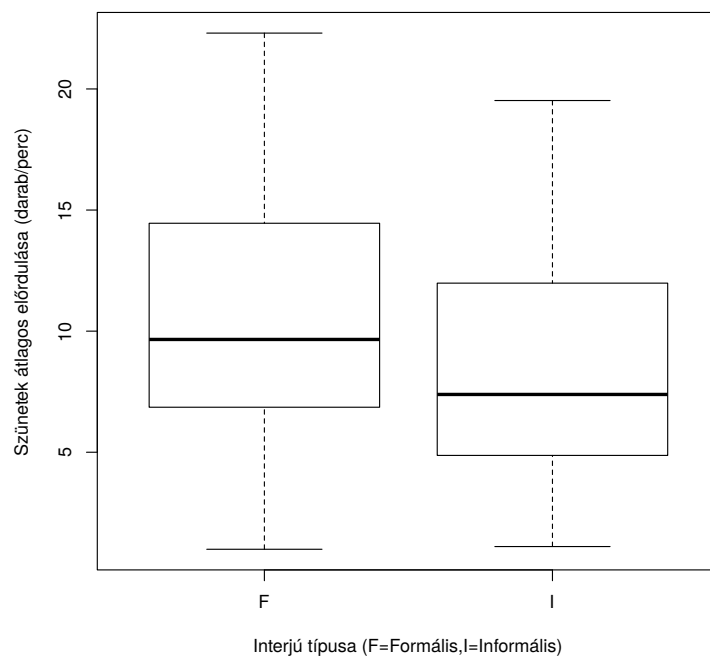
6.1. ábra. Az átfedő beszédek gyakorisága a formális és informális dialógusokban.

viszonyok kiegyenlítődéssel magyarázható. Az informális dialógusokban az interjúalany bizalmának elnyerése, instruálása érdekében az interjúvezető részben maga is adatközlővé változott, míg a szimulált állásinterjúk során szerepe javarészt a kérdések feltevésére és a válaszok megerősítésére korlátozódott.

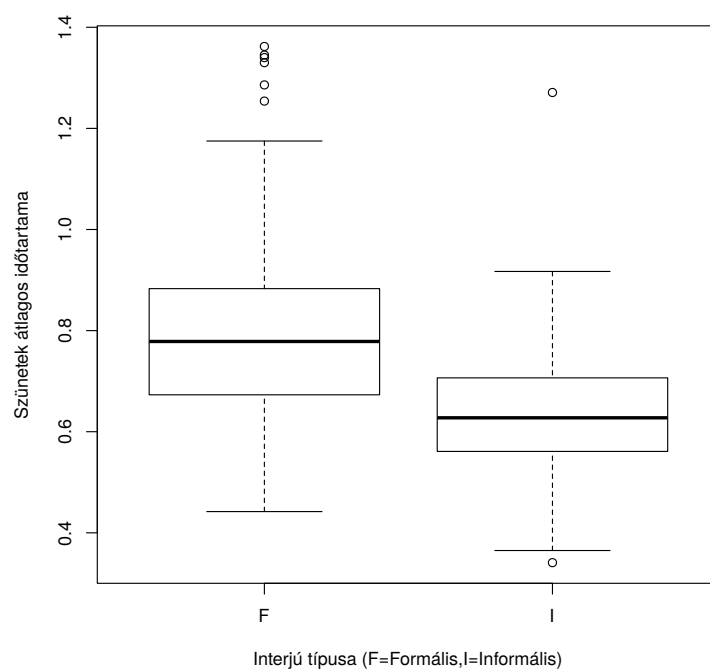


6.2. ábra. Az interjúvezető fordulóinak gyakorisága a formális és informális dialógusokban.

Az előzetes várakozásokhoz képest meglepő eredménynek bizonyult, hogy – habár a formális dialógusokban nagyobb értékek mérhetők – a szünetek gyakoriságában (6.3. ábra) és átlagos időtartamában (6.4.) nem találtunk hasonló mértékű különbséget.



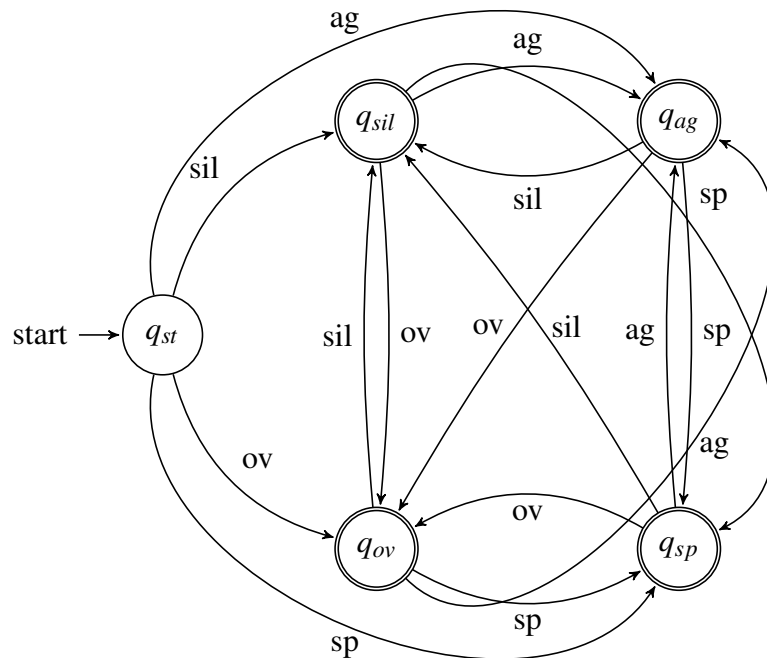
6.3. ábra. A szünetek gyakorisága a formális és informális dialógusokban.



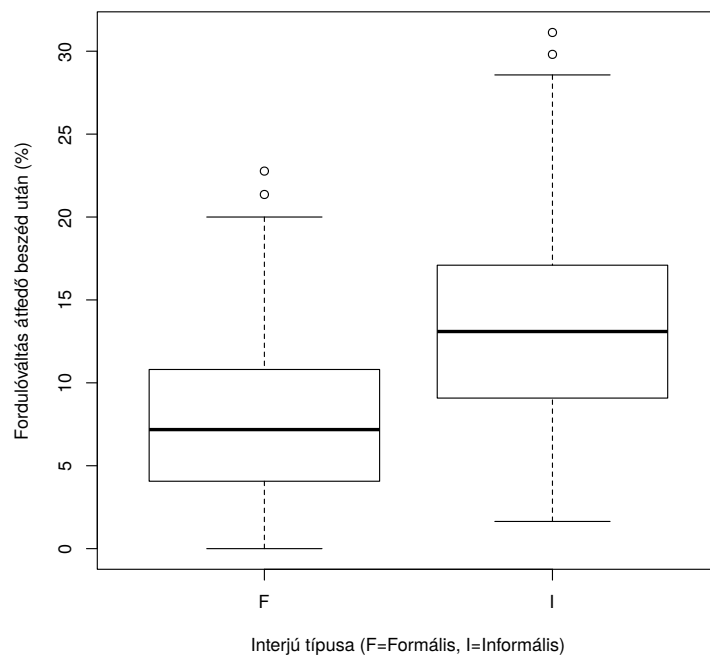
6.4. ábra. A szünetek átlagos időtartama formális és informális dialógusokban.

A beszélőváltással kapcsolatos adatokat a HuComTech korpuszban manuálisan elkészült, a ProsoTool előfeldolgozó algoritmusával (lásd. 4.2.1.3 fejezet) átalakított transzkripciókat felhasználva az egyes fordulókat reprezentáló események szekvenciáit kategorikus változóként bocsátottuk a neuronhálók rendelkezésére. Mint az alábbi példában látható, az időzítésre vonatkozó információkból – az előző szakaszban említett szempontok miatt – csak a fordulók időtartamát és az események sorrendiségét őriztük meg.

sequence	duration
agent	3.799
silence	0.870
speaker	1.594
silence	0.329
agent	2.616
overlap	0.318
speaker	2.532
silence	0.347



6.5. ábra. A fordulóváltások lehetséges mintázatainak ábrázolása véges állapotú automatával. Jelölések: ag=intervjúvezető („agent”), sp=adatközlő („speaker”), sil=szünet („silence”), ov=átfedés („overlapping speech”). A példa transzformációs szabályai akár „üres” (csak szünet), illetve „egyoldalú” (csak az egyik szólal fel) társalgásokat is megengednek.



6.6. ábra. Az átfedő beszédet követő fordulóváltások az összes fordulóhoz viszonyított aránya a formális és informális dialógusokban.

A 6.5. ábra a fordulóváltások lehetséges mintázatait véges állapotú automataként szemlélítve mutat rá a négy kategória (adatközlő beszél, interjúvezető beszél, átfedő beszéd, szünet) kapcsolódási variációinak sokféleségére. Az időtartamok és az egyes kategóriák gyakoriságán túlmenően ezek a mintázatok – amennyiben a két halmaz esetében jellemző eltérést mutatnak – szintén fontos támpontot jelenthetnek a formális vagy az informális halmazba tartozás valószínűsítéséhez. Célszerű lehet például megkülönböztetni azt a környező események által meghatározott esetet, amikor az átfedőbeszédet közvetlenül fordulóváltás is követi (ezt az esetet reprezentálják a fenti példa 5–7. szekvenciái: *agent*, *overlap*, *speaker*), ami a fordulóátvétel egy „erőszakosabb”, de legalábbis a formalitásokat (megvárom, amíg a másik befejezi közlendőjét) kevésbé betartó, kötetlenebb formájára utal. Mint a 6.6. ábrán látható, a HuComTech korpusz formális és az informális dialógusai szignifikáns különbséget (a páros t-próba eredménye: $t = -25.6401$, $df = 335$, $p < 2.2e - 16$) mutatnak abban a tekintetben, hogy az átfedő beszédet követő fordulóváltások milyen arányban fordulnak elő az összes fordulóhoz képest.

6.1.2.2. Intonáció

Az intonáció változásait jellemző adatsorokat kétféle verzióban is előkészítettük a neuronhálók számára. Az első verzió a Praat program alapértelmezett beállításával, 0.01 másodperc hosszú

ablakokra mért alapfrekvencia-adatokat tartalmazta, amelyeket a beszélőváltással kapcsolatos információkkal is kiegészítettünk az alábbi formában:

sequence	f0
agent	188.74473178363115
agent	188.68375296533247
agent	188.37873117477932
agent	188.3452128501004
agent	187.7566988086422
agent	188.62860541647058
agent	188.7505759811598
agent	190.08874412855627

A példában szereplő *sequence* kategorikus változó az aktuális beszélő azonosítóját („agent”), az F0 folytonos változó pedig az adott ablakra mért alapfrekvencia-értéket mutatja. Az egyes fordulók időtartamát itt tehát az azonos beszélőazonosítóval rendelkező sorok száma fejezi ki.

A intonáció reprezentációjának második verziója a ProsoTool kimenetének felhasználásával, a nyers F0-adatok kategorikus változókba (dallamkontúrok és hangmagasságsszintek) történő konverziójával készült el, és szintén kiegészült a beszélőváltással kapcsolatos információkkal:

sequence	duration	countour	level1	level2
overlap	0.325	level	M	M
speaker	2.158	level	M	M
speaker	1.175	fall	M	L2
agent	1.275	level	H1	M
speaker	0.283	rise	L1	H1
speaker	0.433	fall	H1	L1
speaker	0.733	rise	L1	M
overlap	0.212	level	L2	L1

A kísérlet egyik fontos célja pedig annak vizsgálata volt, hogy kettő közül melyik reprezentáció növeli jobban az osztályzás hatékonyságát: a nagyobb adatmennyiséget képviselő, az intonáció legapróbb változásait is feltáró, folytonos F0-adatok, vagy ugyanezek a ProsoTool eljárásával feldolgozott, a kisebb mozgásokat nagyobb trendekben integráló kategorikus minősítései.

6.1.3. A neuronhálók tanítása és tesztelése

A kísérletben használt előrecsatolt mély neuronhálók¹ három, 250 neuronból álló rejtett és egy 2 neuronból álló kimeneti réteget tartalmaztak. Az adatokat pedig a 6.3 táblázatban látható módon osztottuk fel tanító (75%), validációs (10%) és teszt (15%) halmazra.

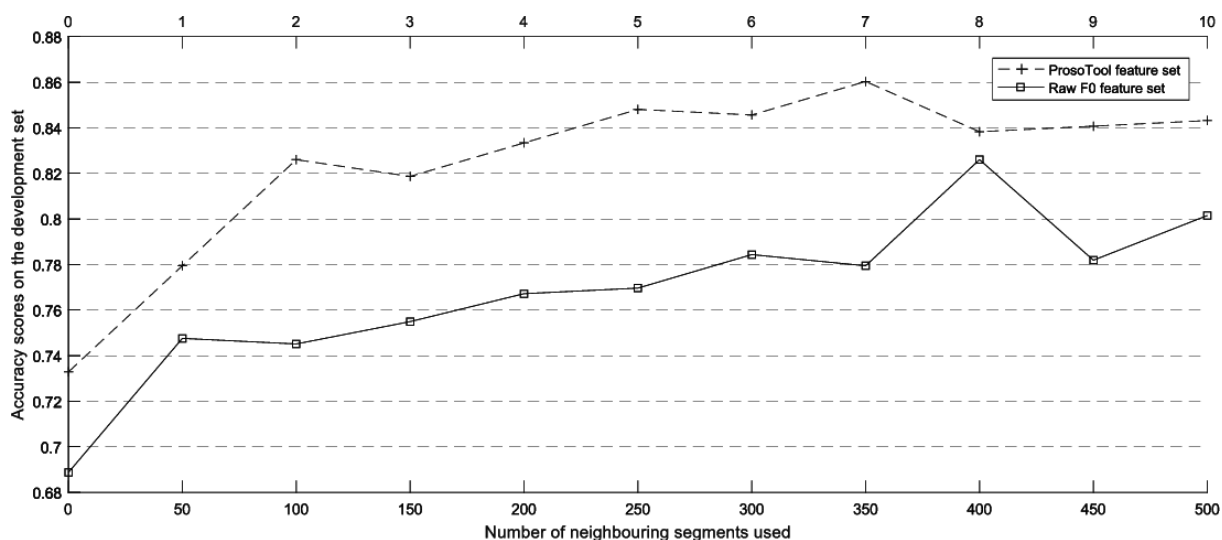
6.3. táblázat. A tanítóanyag felosztása.

	Adatközlők száma	Formális részlet	Informális részlet
Tanító halmaz	83	521	537
Validációs halmaz	11	68	68
Teszt halmaz	17	108	108

A tanítás mind a három jellemzőcsoport (csak beszélőváltás, beszélőváltás + nyers F0, beszélőváltás + ProsoTool kimenet) esetén azonos módszerekkel történt. A háló egy meghatározott méretű ablak (ahol az ablak méretét a benne szereplő szegmensek száma határozza meg) mozgatásával (1 szegmenssel való eltolás) végigjárta az egyes dialógusrészletekhez tartozó adatsorokat, és az ablakban talált szomszédos (balra és jobbra eső) szegmensek kontextusa alapján határozta meg az aktuális szegmens formális vagy informális halmazba való tartozásának valószínűségét. A tesztelés során ezekre az ablakokra kapott valószínűségeknek az átlaga döntötte el az osztályzás kimenetét, amit – mivel a hálók egyszerre csak meghatározott számú szegmenssel találkoztak – a dialógusrészletek eltérő hossza így nem befolyásolt. Mivel a különböző jellemzők esetén az egyes szegmensek eltérő hosszúságú és természetű eseményeket reprezentáltak (beszélőváltás: teljes beszédforduló, F0-adatok: 10 milliszekundum hosszú időablak, ProsoTool: változó hosszúságú dallammenet), a felhasznált kontextus mérete (a szomszédok száma) eltérő léptékben befolyásolta a találati arány változását.

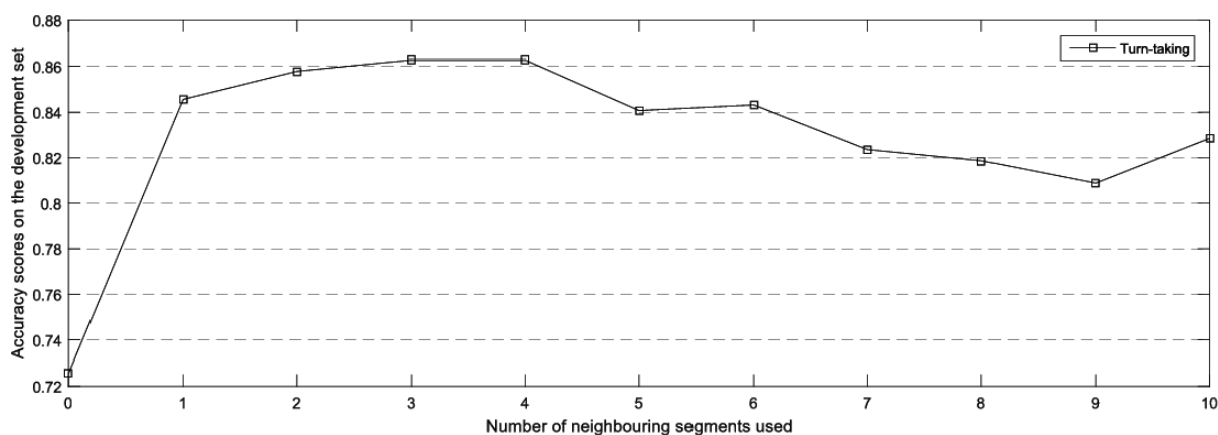
A 6.7. ábra a nyers F0-adatokkal és ProsoTool kimenetével tanított hálók a validációs halmazra kapott találati arányának változását szemlélteti a felhasznált kontextus méretének (F0-adatok: alsó tengely, ProsoTool: felső tengely) függvényében. Mint az ábrán megfigyelhető, a nyers F0-adatokkal dolgozó háló lényegesen nagyobb kontextus figyelembevételével tudott eredményesen működni, mivel a szegmensek időbeli rövidsége miatt csak így nyert úgy mond „rálátást” az intonáció és az egymást követő fordulók mintázataira. 100 szomszédtól 300 szomszédig egyenletes, 4 százalékos kitevő javulás figyelhető meg, ami 400 szomszédnál megugorva éri el a maximális 82.6 százalékos. Ugyanezt az eredményt a ProsoTool jellemzőkkel tanított háló már 2 szomszédos szegmens bevonásával produkálja, 7 szomszédig tovább javul és 86 szá-

¹ A hálók kialakításának és tanításának módszerei elsősorban szerzőtársam (Kovács György) munkáját dicsérik.)



6.7. ábra. A találati arány változása a nyers F0 (lent) és a ProsoTool (fent) jellemzők alapján felhasznált szomszédok szerint (Szekrényes & Kovács 2017).

zalékon éri el a tetőpontot. Az intonációra vonatkozó adatok nélkül tanított háló (lásd 6.8. ábra) eredménye már 1 szomszéd esetén elég magas (84%) és további 3 szomszéd bevonásával érik el a maximális 86.3 százalékot. A tanulás során felhasznált szomszédos szegmentumok számát, avagy az „előre-hátratekintés ablakát” az F0 adatsorokra és a ProsoTool kimenetére támaszkodó hálók esetén is dinamikusan állítottuk be, mégpedig a szerint, hogy a fejlesztői halamazon melyik beállítással volt elérhető pontosabb klasszifikációs eredmény. Mint ez az 6.7. ábrán látható, az F0-adatok esetén egészen 500-500 szomszédig eljutottunk, de nem tapasztaltunk javulást. A már 400 szegmensnél is napokban mérhető tanítási idő viszont még tovább emelkedett. Tekintve, hogy ProsoTool kimenetében szereplő szegmentumok átlagos hossza 700 milliszekundum, 7-7 szegmentum esetén 9800 milliszekundum méretű ablakokról beszélünk, ami tehát egészen jól összemérhető az F0-adatoknál legjobb pontosságot adó 8000 milliszekundummal.



6.8. ábra. A találati arány változása a fordulóváltás alapján felhasznált szomszédok száma szerint (Szekrényes & Kovács 2017).

6.1.4. A kísérlet eredményei

A 6.4. táblázat a validációs halmazon legjobb eredményt elért hálók (beszélőváltás: 4 szomszéd, F0 adatok: 400 szomszéd, ProsoTool: 7 szomszéd) teszhalmazra kapott eredményeit mutatja. A pusztán a beszélőváltásra támaszkodó háló eredményén (81.5%) a nyers F0 adatok 1.1 százalékot rontottak (80.4%), míg a ProsoTool kimenetét felhasználva 3.7 százalékkal bizonyult hatékonyabbnak (85.2%) a formális és informális dialógusrészletek megkülönböztetése.

6.4. táblázat. A különböző jellemzőkkel (nyers F0 adatok, ProsoTool kimenet, fordulóváltás) betanított neuronhálók osztályozási eredményei.

Jellemző set	λ	A felhasznált szomszédos szegmentumok száma	Találati arány	
			Validáció	Teszt
Nyers F0	1.0	400	82.6%	80.4%
Fordulóváltás	0.9	4	86.3%	81.5%
ProsoTool	1.0	7	86.0%	85.2%

Ha az eredményeket összevetjük a 5.4. fejezetben bemutatott percepció kísérlet eredményeivel, ahol az alanyokat azonos kondíciók és tesztanyag használatával állítottuk egy bináris kimenetű klasszifikációs feladat elé, azt látjuk, hogy a percepción alapuló ítéletek a vártnál gyengébb, a gépet látszólag alulteljesítő eredményt mutattak fel. Erre egy lehetséges magyarázatként szolgál, hogy a kísérleti alanyok nem rendelkeztek megfelelő motivációval, és valóban, egy olyan szituációban, ahol az oktató a hallgatóit arra kéri, hogy váltakozó frekvenciájú szinuszhullámok alapján ítéljék meg, hogy amit hallottak, az egy állásinterjú, vagy egy közvetlenebb hangvételű társalgás volt-e, könnyű arra a feltételezésre jutni, hogy a résztvevők kognitív képességeit a feladat helyett inkább a kísérletvezető épelméjűségével kapcsolatos kétségek költötték le. Tegyük fel, hogy nem így történt, de az „azonos kondíciókkal” kapcsolatban akkor is fontos megjegyezni, hogy habár fizikai értelemben valóban a gépi bemenettel teljesen analóg információkra (frekvenciaváltozás, szünetek és beszélőváltások) hagyatkoztak, az ember esetében jobban tekintettel kell lenni az impulzusok fogyaszthatóságára is. Ennek a szempontnak például zenei hangok használata (a szinuszhullámoknál helyett) nem kizárt, hogy inkább megfelelt volna.

A „fairplay” szempontjából további dilemmákat vet fel a „tanulás” kérdése is, hiszen a neuronhálóknak a tesztanyaggal való találkozás előtt, a tanítóhalmazban lévő sok ezer dialógusrészlet alapján lehetőségük volt egy, a két kategóriát jellemző modell felépítésére. A kísérleti alanyok hasonló „képzését” persze szükségtelenné tette az, hogy számukra ugyanez a modell nem csak sok ezer példa, de összemérhetetlen mennyiségű tapasztalat alapján eleve adottnak (és a gép tanításában is mérvadó mintának) tekinthető. Az ember számára azonban ezek, a társalgás

különböző típusairól alkotott fogalmak feltehetően egészen más formában képezik részét egy a gép által felépített modellnél sokkal bonyolultabb szemantikai szisztémának, amelyben elsősorban nem a tapasztalt példák prozódiai sajátosságai alapján, hanem egyéb, magasabb prioritást élvező funkcionális szempontoknak megfelelően integrálódtak. Kísérletünkben a „gépi agy” eleve a választott szempontoknak megfelelően igyekezett disztinkciót tenni a két kategória között, ami a HuComTech korpusz zárt rendszerében relatíve sikerre vezetett, de ez nem feltétlen jelenti azt, hogy ezek a szempontok minden egyéb esetben relevánsak, a kategóriákat elégségesen vagy szükségesen jellemző attribútumok volnának az emberi kogníció számára is, ha egyáltalán képesek vagyunk tudatosan reflektálni a gép által könnyebben kezelhető, az interakciók időszerkezetét érintő mintázatokra.

6.2. A témaváltások detektálása

Egy másik – szintén mély neuronhálók segítségével implementált – kísérlet célja a HuComTech korpusz diskurzusaiban található témaváltások automatikus detektálása volt, amelyhez jellemzőként az előzőhöz hasonlóan a fordulóváltásokkal kapcsolatos adatok és a ProsoTool kimenete került felhasználásra. A témaváltás detektálása a szakirodalomban korántsem előzmény nélküli terület. Főként szövegek automatikus szegmentálására történtek lexikai alapokon implementált kísérletek (Kozima, 1993; Reynar, 1994; Sitbon & Bellot, 2007), de találhatunk olyan hangzó anyagok feldolgozását célzó példákat is, amelyek a transzkripciókban lévő szöveges információkat kiegészítettek a megnyilatkozások prozódiai jellemzőivel is (Passeau & Litman, 1997; Molugu, 2003), vagy tisztán akusztikai információkra támaszkodtak (Hirschberg & Nakatani, 1996, 1998). Külön kiemelném Luz & Su (2010) munkáját, akik a témaszerkezet szegmentálását a beszélőváltás szünetekkel kiegészített szerkezte alapján az AMI korpuszon² (Carletta et al., 2006) végezték el.

Saját kísérletünkben a HuComTech korpusz dialógusainak az interjúvezető által feltett, előre eltervezett kérdések mentén történő automatikus szegmentálására törekedtünk. Ez lényeges változtatást jelentett Kovács & Váradi (2017) azon korábbi munkájához képest, amiben hasonló céllal, de a korpusz pragmatikai annotációban jelölt, a témaszerkezet szempontjából releváns kommunikációs aktusok (*topic initiation*, *topic change* és *topic elaboration*,) alapján történt a neurális hálózatok betanítása és nem kizárólag a prozódia, de a korpuszban – annotáció formájá-

² <http://groups.inf.ed.ac.uk/ami/corpus/>

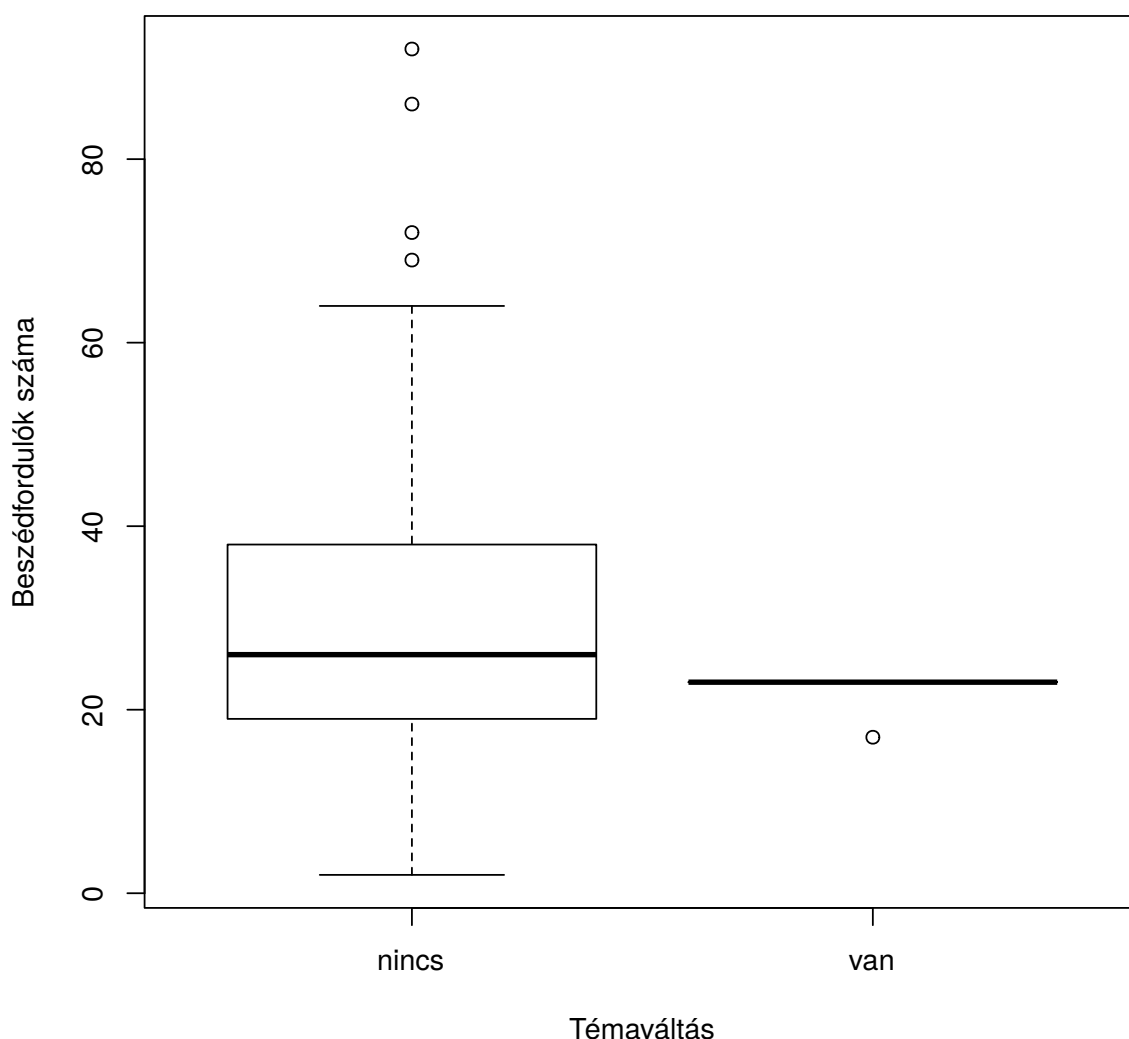
ban – felelhető összes jellemző felhasználásával.³ A változtatás oka az volt, hogy a pragmatikai annotációban az annotációt végző személyek szubjektív döntései alapján jelölt típusok magukba foglalták azokat az eseteket is, amikor váltás egy adott interjú témán belül maradván (akár az interjúalany kezdeményezésére), a társalgás természetes folyamatát követve, motivált módon következett be. Az új kísérletben kizárólag az előre eltervezett, a társalgás menetébe „erőszakkosan”, az interjúvezető „hatalmi” aktusa által bekövetkező témaváltásokra fókuszáltunk, amelyek – hipotézisem szerint – nagyobb valószínűséggel hagyhatnak nyomot az interakció prozódiai szerkezetében is. A témaváltások detektálása szempontjából természetesen előnyt jelentett, hogy ezek az aktusok kizárólag az interjúvezetőhöz kötődtek, így a beszélőváltás szerkezete még hasznosabb jellemzőnek bizonyult, mint az előző kísérletben.

Mint a 6.9. és a 6.10. ábrán látható, a formális és az informális interjúk között jelentős különbség van a tekintetben, hogy a interjúvezető beszédfordulói milyen arányban járnak együtt a téma megváltozásával. Míg a formális interjúkban az interjúvezető megszólása (AS) már önmagában is a témaváltás (TS) közelítőleg 50 százalékos esélyét vonja magát, addig az informális dialógusokban – az interjúvezető fokozott aktivitása miatt – ez sokkal kevésbé garantált: $P(TS|AS) < 0.1$.

Egy másik fontos változtatás a prozódiai jellemzők előző fejezetek bemutatott reprezentálását illetően történt. Mivel a témaváltás detektálása a dialógusrészletek osztályzásánál (formális vagy informális) sokkal komplexebb, az interakciót jellemző prozódiai tulajdonságok időbeli változásaira érzékeny feladat, a hangmagasságsszintek kategorikus ábrázolását (L_2 , L_1 , M stb.) nem éreztük kellően részletesnek a neurális hálók eredményes betanításához. Ezért kategorikus változók helyett a ProsoTool kimenetében szereplő stilizált F0-értékek normalizált verzióját használtuk fel. A normalizálás szerepe itt is az volt, hogy a konkrét értékeket a beszélő egyéni – az alapfrekvencia átlagával és szórásával kifejezett – beszéd-sajátosságaihoz viszonyítsuk. A különbség csak annyi, hogy a normalizálás eredménye nem egy, a relatív különbségeket megragadó kategória (hangmagasságsszint) lett, hanem egy folytonos változóként kezelhető, standardizált érték, ahol a zérus fejezi ki az eredeti, a stilizált dallamszegmentumok elején és végén mért értékek átlagát:

$$z_i = \frac{x_i - \bar{x}}{\sigma} \quad (6.1)$$

³ A prozódia alapján a többihez képest is elég alacsony (51,5% UAR) eredmény volt csak elérhető: a 6 jellemzőcsoportból a prozódia harmadik helyen végezett.

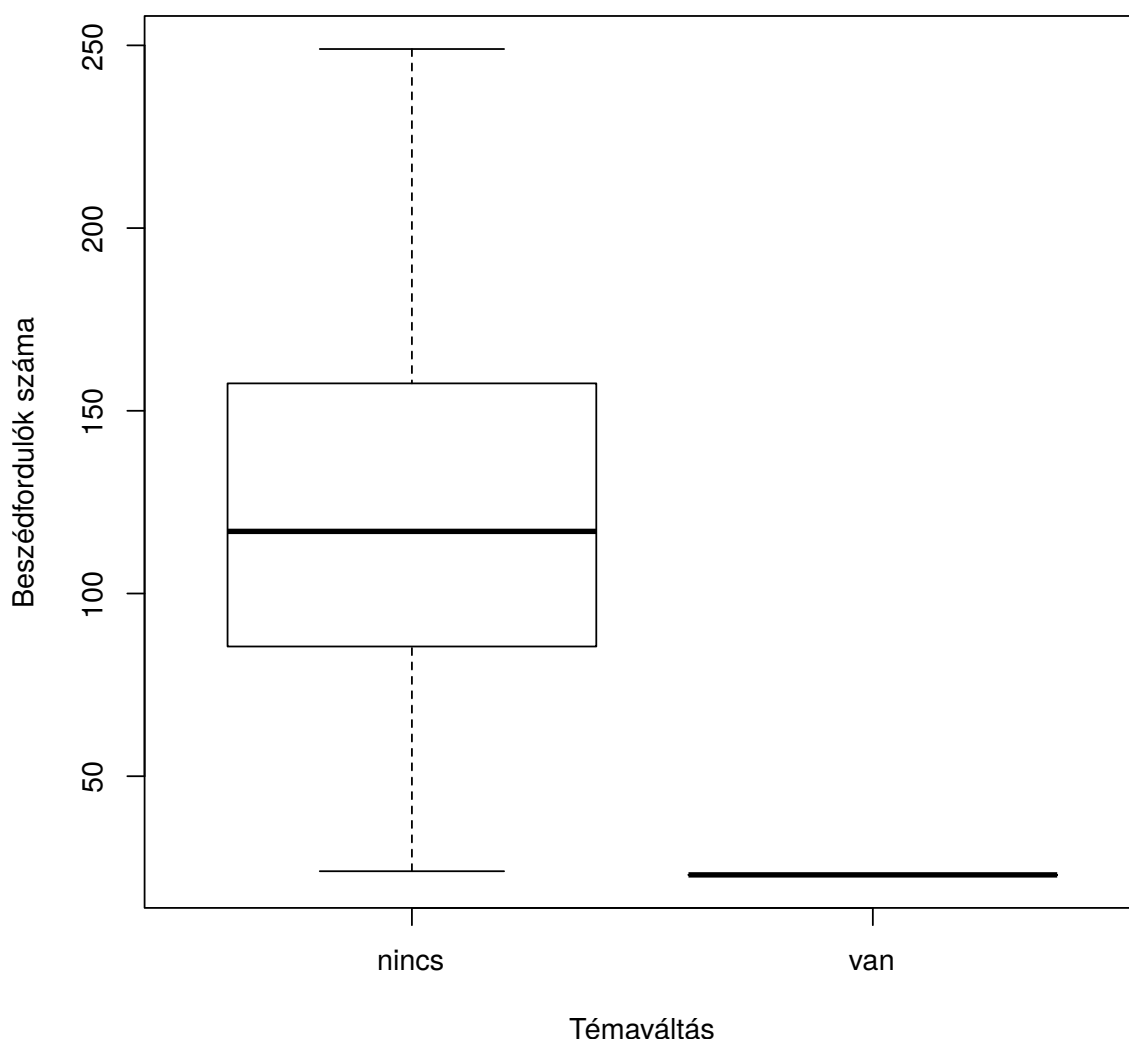


6.9. ábra. Az interjúvezető témaváltással járó beszédfordulóinak aránya a formális dialógusokban.

A dallamszegmentumok kontúrját (a kategorikus reprezentációban: „rise”, „fall” stb.) a két érték közötti különbözetként jellemeztük, ami szintén normalizálásra került azzal a különbséggel, hogy a zérus itt nem a különbségek átlagát, hanem ténylegesen azt az esetet fejezi ki, amikor a hangmagasság tekintetében nincs változás:

$$z_i = \frac{x_i - \bar{x}}{(\sigma - \frac{0 - \bar{x}}{\sigma})} \quad (6.2)$$

Az intonációra vonatkozó adatokat kiegészítettük a beszédintenzitás és a szótagfrekvencia alapján mért beszédtempó (lásd 4.2.3. fejezet) teljes szegmentumra vonatkozó átlagával. Természetesen ezeket az értékeket is normalizáltuk a 6.1. képlet használatával.



6.10. ábra. Az interjúvezető témaváltással járó beszédfordulóinak aránya az informális dialógusokban.

A 6.5. táblázat egy példán keresztül szemlélteti, hogyan történt az egyes jellemzők reprezentációja. Az oszlopok tartalma balról jobbra haladva: a téma azonosítója (a tanítás során nem használtuk, csak a könnyebb értelmezés kedvéért szerepel a táblázatban), a fordulóváltás szekvenciái, a szegmentumok időtartama, intonációs kontúr (a hangmagasság változásának mértéke és iránya), a szegmentum elején és végén mért relatív hangmagasságok („level1” „level2”), a szegmentumban mért átlagos beszédintenzitás, a szegmentumban mért átlagos beszédtempó. Az utolsó („tp.shift”) oszlopban bináris formában (van váltás / nincs váltás) emeltük ki azokat szegmentumokat, amelyek egy témaváltással járó beszédforduló részeit képezik. A szünetek helyén az összes prosódiai jellemző esetén egy kiugróan negatív szám (-10000) szerepel, mivel a zérus a többi esetben az átlagos hangmagasság, hangerő és tempó kifejezésére szolgált.

6.5. táblázat. A neuronhálók tanítása során használt jellemzők reprezentációja.

topic	speaker	duration	contour	level1	level2	int.	sp.rate	tp.shift
T0	agent	0.581	-0.133	0.908	0.727	-0.394	-0.957	0
T0	silence	1.191	-10000	-10000	-10000	-10000	-10000	0
T0	speaker	0.392	0.121	0.439	0.666	-0.325	-0.757	0
T0	overlap	0.175	-0.368	2.018	1.617	-0.923	-1.344	0
T0	agent	0.769	-0.653	0.599	-0.288	-0.738	-0.541	0
T0	speaker	0.808	0.015	-0.166	-0.139	-0.424	-0.802	0
T0	agent	0.281	-0.229	0.637	0.325	-0.434	-0.087	1
T0	agent	0.3	-1.279	0.325	-1.415	-0.477	-0.195	1
T0	agent	1.059	1.713	-1.415	0.916	-1.110	-1.326	1
T0	agent	0.562	-1.396	0.916	-0.984	-0.484	0.751	1
T1	agent	1.022	1.514	-0.222	1.837	-2.163	-1.307	1
T1	agent	0.431	-1.050	1.837	0.408	-0.124	-0.673	1
T1	agent	1.95	-0.432	0.408	-0.179	-0.224	-0.560	1
T1	speaker	1.575	0.662	-0.421	0.819	-0.108	-0.763	0
T1	speaker	1.5	-0.689	0.819	-0.471	-0.373	0.045	0
T1	speaker	0.867	0.760	-0.471	0.954	-0.693	-0.254	0
T1	speaker	0.558	-0.573	0.954	-0.121	-0.762	-1.178	0
T1	speaker	0.767	1.517	-0.121	2.723	-0.013	0.720	0

A rekurens neuronhálók tanítása a korpusz Kovács & Váradi (2017) korábbi felosztási arányát követve (75/10/15), az előző fejezetben bemutatott kísérlethez hasonló módon, szegmentumról szegmentumra haladva, a szomszédos szegmentumok figyelembevételével zajlott. Volt azonban néhány fontos különbség. Egyrészt, hogy a különböző jellemzőcsoportok (beszélőváltás és prozódia) kombinációjának hatékonyságát egymástól izoláltan tanított hálók fúziójával is vizsgáltuk, másrészt a szegmentumokra kapott valószínűségeket itt nem a dialógusrészletekre, hanem az egyes beszédfordulókra aggregáltuk.⁴ A Kovács György által implementált metódusok közül az alábbiakban csak a beszélőváltás és a prozódia alapján (egymástól izoláltan) tanított hálók, illetve ezek fúziójából származó eredményeket fogom ismertetni.

A prozódiai jellemzőkön tanított hálók („PROSODY-ONLY”) első lépésben a szegmentumok szintjén végezték a klasszifikációt, ami a beszédfordulókra való aggregáláskor bemenetként megkapta a forduló időtartamát, illetve a forduló egészére átlagolt beszédtempó normalizált értékét is. Ezzel párhuzamosan egy másik háló a fordulók szintjén, csak a beszélőváltás adataira és a forduló időtartamára hagyatkozott („TURN-TAKING”). A két háló döntésének fúziójával („LATE-FUSION”) pedig jobb eredményeket értünk el, mint amikor egy harmadik háló már a szegmentumok szintjén vegyesen támaszkodott a két jellemzőcsoportból származó információkra („EARLY FUSION”).

⁴ Hiszen az osztályzás célja a fordulók kategorizálása volt a dialógusok interjútemákra való szegmentálása érdekében.

A tesztalmazra kapott eredményeket az F_1 mérték metrikája (lásd 6.5. képlet) alapján értékeltük, ami esetünkben a tényleges témaváltások (TTV), a helyesen felismert témaváltások ($HFTV$) és az összes témaváltásként felismert fordulók (FTV) számának figyelembevételével az alábbi módon volt implementálható:

$$precision = \frac{HFTV}{FTV} \quad (6.3)$$

$$recall = \frac{HFTV}{TTV} \quad (6.4)$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (6.5)$$

6.6. táblázat. A témaváltás detektálásával kapcsolatos kutatások eredményeinek összehasonlítása (a jelen fejezetben ismertetett kísérlethez tartozó F_1 mérték vastagon kiemelve).

Módszer	Jellemzők	F_1 mérték
Luz and Su (2010)	beszélőváltás	0.28
Molugu (2003)	prozódia + szöveg	0.33
PROSODY-ONLY (Kovács & Szekrényes, 2018)	prozódia	0.36
Hirschberg and Nakatani (1998)	prozódia	0.41
TURN-TAKING (Kovács & Szekrényes, 2018)	beszélőváltás	0.46
Passoneau and Litman (1997)	prozódia + szöveg	0.48
LATE FUSION (Kovács & Szekrényes, 2018)	prozódia + beszélőváltás	0.49

A 4.3. táblázat a hasonló nemzetközi kutatásokkal összehasonlítva mutatja be a kísérlet eredményeit. A táblázatban látható, hogy habár a kizárólag prozódiai jellemzőket (a ProsoTool normalizált kimenete) használó módszerünk Hirschberg & Nakatani (1998) kísérletéhez képest néhány ponttal lemaradt, az F_1 mérték tekintetében még mindig összemérhető az azonos jellemzőkön vagy beszélőváltásokon alapuló, illetve Molugu (2003) esetén textuális információkat is integráló módszerekkel. A beszélőváltás a várakozásoknak megfelelően hasznos jellemzőnek bizonyult, és meg is haladta a prozódiai alapján történt predikciók eredményét. A kettő fúziójával pedig a felsoroltak közötti legpontosabb detektálást sikerült kivitelezni. Az összehasonlítás természetesen problematikus abból a szempontból, hogy különböző algoritmusok tesztelése nem azonos tesztanyagon történt, illetve a témaváltás meghatározása esetünkben egy szűkebb, önkényesen választott típusra, a „motiválatlan” interjúkérdések esetére korlátozódott.

7. FEJEZET

Összefoglalás

Az értekezést az eredmények rövid összegzésével, a legfontosabb konklúziók tézisként való megfogalmazásával zárom. A dolgozat szövegében a kutatott téma tárgyalását három nagyobb egységre csoportosítottuk: módszertani fejlesztések (4. fejezet), percepció, illetve gépi tanulás-sal végzett kísérletek (5. és 6. fejezet). Mivel az intonáció automatikus annotálására kidolgozott módszer nem csupán a saját hipotéziseim igazolását szolgálta, de eszközként más kutatásokban is hasznosítható lehet, a vonatkozó eredményeket itt is külön kiemelve mutatom be. Egyéb tekintetben viszont nem a disszertáció szerkezeti csoportosítását követem, hanem az egymással való összefüggésük szerint ismertetem a különböző részeredményeket.

Az e-magyar projekt keretében közzétett, a 4. fejezetben bemutatott annotációs eljárás¹ célja az volt, hogy lehetővé tegye az előbeszédben megfigyelhető intonációs jelenségek automatikus címkézését. Az algoritmust a Praat program beépített szkriptnyelvén implementáltam. Használata előfeltételezi, hogy az elemezni kívánt társalgásmintában a beszédfordulók valamilyen módon annotálva vannak, de az átfedő beszédek és az adatközlők hangjának izolálása már az automatikus előfeldolgozás részét képezi. Az eredményül kapott annotáció, a beszélők hangjában megfigyelhető, az alaphangmagasság-értékek simításával és stilizálásával rekonstruált intonációs eseményeket a dallammenet kontúrájára és a relatív hangmagasság változására vonatkozó kategorikus címkék formájában jegyzi le. Ezt a Praat TextGrid formátumú szöveges kimenetet egy webes környezetben kezelhető, XML formátumú verzióval és a stilizált dallamkontúrokat SVG grafikaként megjelenítő XSL stíluslappal is kiegészítettem, amit később az eljárás validálására használtam fel. Az értékelést végző alanyok a dallamkontúrok vizualizációit vetették össze a felvételen hallható intonációval. A 25 megnyilatkozásra eddig 9 alanytól kapott értékelések átlageredménye alapján az annotációs eljárás eredménye 82,9 százalékban volt összhangban az

¹ <https://e-magyar.hu/hu/speechmodules/empros>

értékelők percepciójában megjelenő hangélménnyel. A vizualizációk másik, a módszer validáláson túlmutató, tudománynépszerűsítő funkciója, hogy az előbeszéd prozódiajának leképezését a szűkebb kutató közönségen kívül a laikus érdeklődők számára is élvezhető formában mutassa be. Távlati terveim között szerepel, hogy *github* repozitóriumban jelenleg is elérhető Praat szkriptek és XSL sablonok letöltése helyett a prozódiai címkék generálása és vizualizálása közvetlenül az e-magyar projekt weboldalán, egy tetszőleges hangfájl és a hozzá tartozó szöveges átirat feltöltése után kipróbálható legyen. A HuComTech korpusz felvételeire kapott kimenet a korpusz többi annotációjával (szöveges átiratok, egyéb non-verbális gesztusok, pragmatikai és szintaktikai annotáció) együtt az MTA Nyelvtudományi Intézetének infrastruktúrális hozzájárulásával² tettük elérhetővé az érdeklődő kutatók számára. Az ANNEX kezelőfelületén vagy az ELAN programban végrehajtható lekérdezéseken (pl. a különböző prozódiai címkék milyen más gesztusokkal és kommunikációs eseményekkel együtt fordulnak elő) kívül itt fontosnak látom felhívni a figyelmet egy alternatív, a 4.1.2. fejezetben már hivatkozott, az eseményszekvenciák rejtett időbeli mintázatait feltáró THEME program nyújtotta elemzési lehetőségre. Habár az értekezés keretében nem került tárgyalásra, az egyik friss fejlesztésem eredményeként (Szekrényes, 2019) közétett interfész³ ezeknek a mintáknak a feldolgozását könnyíti meg.

A percepciók kísérletek közül a 5.3. fejezetben bemutatott, a hanglejtés disztinktív erejét mérő teszt („Angéla néni” esete) a stilizálás eljárása kapcsán részben a fent leírt annotációs módszer ismételt validálása célján került kidolgozásra. Ebből a szempontból a kísérlet legfontosabb eredményének azt tartom, hogy a kísérleti alanyok – a két kondícióban elért pontok elemzése alapján – ugyanolyan hatékonysággal tudták azonosítani a megnyilatkozásnak leginkább megfelelő kontextust, mint a nyers F0-adatok alapján generált stimulusok esetén, így bár csak 11 résztvevő adatai szerint, de feltételezhetjük, hogy a megnyilatkozások attitűdtartalmának megkülönböztetése tekintetében az alaphangmagasság görbék stilizálásával az eredeti görbékhez képest a percepció nem veszített releváns információkat. Szintén fontosnak látom kiemelni, hogy a „féleső” és az „eső” karakter (az attitűd tartalom tekintetében „önálló”, illetve „önálló és befejezett” kategóriába tartozó megnyilatkozások) elkülönítése a beszélőtől prezentált kis számú minta ellenére is sikeres volt, ami igazolni látszik Varga Lászlónak a két karaktert elkülönítő relatív záróhangmagasságra vonatkozó téziseit⁴, illetve a módszertani megfontolást, ami

² <https://tla.nytud.hu>

³ <https://altnyelv.unideb.hu/ThemeToMySQL>

⁴ „A hallgató általában rövid idő alatt «beméri» a beszélő alapvonalát, és meg tudja állapítani, hogy a beszélő elérte-e azt vagy sem. Ezért az eső és félig eső karakterek különbségét kategorikus különbségnek tartjuk.” (Varga, 1994. 477. o.)

alapján ezekről a relatív hangmagasságokról a prozódiai annotáció kimenetében is számot adok a kontúrok kategóriái mellett.

A beszéd stilizált dallamának kategorikus változókkal való reprezentálására kidolgozott annotáló eljárás egy, a 6.1. fejezetben bemutatott kísérlet során jól felhasználhatónak bizonyult a beszédszituáció („formális vagy informális”) neurális hálózatokkal történő automatikus osztályozásához. Habár a hanglejtéssel kapcsolatos adatok csak néhány százalékos javulást jelentettek a formális és informális interjúk megkülönböztetésében, hasonló javulást a nyers, címkézetlen F0-adatokkal nem tudtunk elérni. A disszertáció egyik nem várt, de fontos eredményének tekinthető, hogy a beszélőváltás szekvenciáinak ritmusát kifejező jellemzőkön tanított hálók már önmagukban is meglepően eredményesek működtek. A 108 dialógusrészletet tartalmazó teszt-halmaz osztályozása a beszélőváltás alapján 81,5 százalékos pontossággal zárult, ami további 3,7 százalékot javult (ami 20% relatív hibaarány csökkenést jelent), ha a jellemzőkészletet a prozódiai annotáció címkéivel (dallamkontúrok és relatív hangmagasságsszintek) is kiegészítettük. Ezzel szemben a nyers F0-adatok bevonása az eredmény 1,1 százalékkal történő romlásához és jelentősen hosszabb (órák helyett napokban mérhető) tanítási időhöz vezetett.

A témaváltások prozódiai jellemzők felhasználásával történő automatikus detektálása az értekezés keretében folyó kutatások egyik másik fontos gyakorlati, de az alaphipotézis szempontjából sem érdektelen célkitűzését jelentette. Az ezzel kapcsolatos, neurális hálózatokkal végzett, 6.2. fejezetben bemutatott kísérlet szintén sikeresen, a hasonló nemzetközi kutatásokkal összemérhető eredményekkel zárult. A ProsoTool algoritmusával generált jellemzőket itt a beszédtempó és a beszédintenzitás értékeivel egészítettük ki. A hálók tanításánál a korábbi módszer elveit megőrizve nem csak az aktuális megnyilatkozás prozódiai sajátosságait, de annak környezetét is figyelembe vettük.

A 5.1. fejezetben elsőként bemutatott kísérletek esetében a stimulusok kis mennyisége, a függő változók kontrollálásával kapcsolatos tervezési problémák, illetve a feladat nehézsége miatt nehéz határozott, számszerűsíthető eredményeket felmutatni.⁵ Az itt vizsgált, a prozodiához köthető, alul-specifikált információtartalom relevanciájával kapcsolatos hipotézist sem megerősíteni, sem cáfolni nem tudtam, ugyanakkor a kapott adatok alapján megalapozottnak látom a tesztek gondosabban kontrollált kondíciók és tudatosabban összeválogatott stimulusok mellett történő folytatását. A 2.3. fejezetben felvetett al-hipotézist, miszerint az interjú-szituációban lezajló témaváltások markánsabb nyomott hagynak az interakció prozódiai szerkezetében, né-

⁵ A résztvevők számát illetően ez a teszt pont kivételt képez, de sajnos a többi (a koncentráció képességét változó mértékben terhelő) percepciók kísérleteimre általában igaz, hogy a normákhoz képest kevés alany közreműködésével tudtam lebonyolítani őket.

mileg erősíteni látszik az a tény, hogy a teszt során használt négy felvétel közül csak ebben az egy speciális esetben voltak többségben azok, akik a témaváltást a verbális tartalom nélkül is egyértelműen érzékelték.

Érdekes kérdéseket vet fel a formális és informális dialógusrészletek megkülönböztetését célzó, 5.4. fejezetben bemutatott percepciós kísérlet negatív eredménye is (a többségi döntés pontossága csak 56% volt). Az azonos kondíciók mellett betanított neurális hálózatok teszteredményeivel összevetve úgy tűnik, hogy a percepción alapuló ítéletek mintegy alulteljesítették a géppel végzett klasszifikációt. Az adott fejezetben ezt két tényezővel magyaráztuk. Egyrészt a társalgás különböző típusairól alkotott fogalmak az ember estében feltehetően egészen más formában képezik részét egy a gép által felépített modellnél sokkal bonyolultabb szemantikai szisztémának, amelyben elsősorban nem a tapasztalt példák prozódiai sajátosságai alapján, hanem egyéb, magasabb prioritást élvező funkcionális szempontoknak megfelelően integrálódtak. Az, hogy a neurális háló az általam választott jellemzők segítségével sikeresen tudott disztinkciót tenni a két kategória között, nem feltétlen jelenti egyúttal, hogy a kategóriákat elégségesen és szükségesen jellemző attribútumok volnának az emberi kogníció számára is. Másrészt a használt stimulusok habár fizikailag valóban ekvivalensek voltak, a szinuszhullámok hallgatása az alanyoknál nem biztos, hogy egyenértékű támpontot adott azoknak a képességeknek, amelyek a természetes beszéd esetén lépnek működésbe. A 5.2. fejezetben, a prozódiai jellemzők követhetőségét mérő kísérletben ugyebár azt láttuk, hogy a beszédből kinyert mesterséges jel megkönnyíti a változásokra való koncentrációt, ellenben ha a felvételeken izolálható beszédhangokat és szavakat hallottak, akkor a percepcióban ösztönös módon a prozódiai jellemzők is a beszélők megnyilatkozásainak és attitűdjeinek értelmezését segítő nyelvi, illetve kommunikatív funkciója felől nyertek megvilágítást. A formális és informális társalgások megkülönböztetéséhez talán éppen ez hiányzott, ugyanakkor az „Angéla néni” megnyilatkozás (lásd. 5.3. fejezet) szinuszhangként szintetizált verzióinak értelmezésekor nem jelentett akadályt.

A főt megfogalmazott megfontolásokkal a gépi–emberi teljesítmény összevetésének legitimitásával kapcsolatos kétségeimnek kívántam hangot adni, amivel nem célom a gépi tanulással elért eredmények jelentőségét csökkenteni. Habár köztudott, hogy mesterséges intelligenciával kapcsolatos kutatásokban nagy kihívást jelent az ember számára rutinfeladatot jelentő műveletek (pl. a beszéd felismerése, autóvezetés, járás közbeni egyensúlyozás) implementálása, vannak olyan jelenségek, amelyek esetében nem az emberi kogníció reprodukálása (netalán kiváltása) a cél, hanem azok sajátunktól eltérő perspektívában való megvilágítása. Úgy gondolom, hogy a disszertáció keretében vizsgált, a társalgás „makro-prozódiját” képező alakzatok, amelyeket az interakcióban részt vevő személyek megnyilatkozásaik ösztönös megformálásával és

pozicionálásával alakítanak ki, ha a humán megfigyelő számára néha rejtve is maradnak, ez nem jelenti azt, hogy a percepcióra nincs hatásuk, és hogy más módszerekkel ne lehetne kimutatni, vagy géppel történő elemzés után hasznosítani őket.

A disszertáció keretében folytatott kutatás eddigi eredményei alapján levont következtetések az alábbi tézisek formájában foglalhatók össze:

1. Az előbeszéddel kapcsolatos percepció és információtartalom tekintetében releváns prozódiai változások szabály alapú algoritmusokkal való felcímkézése (stilizálás, relatív alapfrekvencia-viszonyok vizsgálata) nem csak lényegesen gyorsabb, mint a humán annotációt és gépi tanulást alkalmazó eljárások, de az eredmény minőségében is összemérhető velük.
2. Kimutattuk, hogy egy deskriptív, a kontúrok elemzésében fonológiai szempontokat nélkülöző, szabály alapú annotációs modell magyar nyelvű anyagon alkalmas az érzékelhető prozódiai változások számítógépes feldolgozására. Mivel a módszer nyelv-specifikus elemeket nem tartalmaz, lehetségesnek tartom, hogy egyéb nyelvekre is alkalmazható.
3. Az alapfrekvencia-kontúr megfelelő stilizálása a percepció vonatkozásában nem eredményez szignifikáns információvesztést a különböző hanglejtésformák által kifejezésre jutó attitűdtartalmak sikeres azonosításában.
4. A nyers akusztikai jellemzők szabály-alapú feldolgozása és kategorikus változók formájában történő reprezentációja a neurális hálózatok tanulási idejét hatékonyan csökkenti és nem eredményez olyan információvesztést, ami az előszóban zajló társas interakciók a beszédhelyzet (formális vagy informális) szempontjából történő osztályozását rontaná.
5. A beszélőváltás szünetekkel és átfedőbeszédekkel, illetve prozódiai információkkal is kiegészített szekvenciáinak ritmusa a verbális tartalom hiányában is hatékony jellemzőként használható fel a kommunikációs partnerek viszonyával, szerepével és a beszédhelyzettel kapcsolatos globális tulajdonságok meghatározásához.
6. A témaváltás verbális tartalomtól független, pusztán a prozódiai szerkezetre támaszkodó detektálása motiválatlan témaváltások esetén hatékonyan implementálható, ha mély neuronhálók tanításánál és az osztályzás során nem csupán az új témát bevezető megnyilatkozást vizsgáljuk (kezdeti F0-érték, hangterjedelem), hanem az azok környezetében előforduló hanglejtés szekvenciák tulajdonságait és mintázatait is.

Irodalomjegyzék

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Retrieved from <https://www.tensorflow.org/> (Software available from tensorflow.org)
- Abuczki, A. (2013). A mondjuk nem konceptuális használatának vizsgálata multimodális kontextusban. In T. Váradi (Ed.), *Alknyelvdok7* (pp. 303–316). Budapest: MTA Nyelvtudományi Intézet.
- Abuczki, A. (2014). *A core/periphery approach to the functional spectrum of discourse markers in multimodal context* (PhD dissertation). Debreceni Egyetem, Nyelvtudományok Doktori Iskola.
- Anguera, X., Wooters, C., & Pardo, J. M. (2006). Robust speaker diarization for meetings: ICSI RT06s meetings evaluation system. In *Machine learning for multimodal interaction* (pp. 346–358). Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/11965152_31 doi: 10.1007/11965152_31
- Austin, J. L. (1962). *How to do things with words*. Cambridge: Harvard University Press.
- Baiat, G. E., & Szekrényes, I. (2012, Dec). Topic change detection based on prosodic cues in unimodal setting. In *2012 IEEE 3rd International conference on cognitive infocommunications (CogInfoCom)* (p. 527-530). doi: 10.1109/CogInfoCom.2012.6422037
- Beke, A. (2015). *Gépi beszélődetektálás magyar nyelvű spontán társalgásokban*. Budapest: ELTE Eötvös Kiadó.
- Berck, P., & Russel, A. (2006). ANNEX – a web-based framework for exploiting annotated media resources. In *Proceedings of LREC 2006* (pp. 1552–1555).
- Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *IFA proceedings 17* (pp. 97–110).

- Boersma, P., & Weenink, D. (2016). *Praat: doing phonetics by computer [Computer program]. version 6.0.22*. <http://www.praat.org/>. (retrieved 15 November 2016)
- Bolinger, D. (1972). Introduction. In D. Bolinger (Ed.), *Intonation. Selected readings*. (pp. 11–15). Harmondsworth–Baltimore–Ringwood: Penguin Books.
- Bonastre, J. ., Wils, F., & Meignier, S. (2005, March). ALIZE, a free toolkit for speaker recognition. In *Proceedings of (ICASSP '05). IEEE international conference on acoustics, speech, and signal processing, 2005*. (Vol. 1, p. I/737-I/740 Vol. 1). doi: 10.1109/ICASSP.2005.1415219
- Broeder, D., Claus, A., Offenga, F., Skiba, R., Trilsbeek, P., & Wittenburg, P. (2006). LAMUS: The language archive management and upload system. In *Proceedings of LREC 2006* (pp. 2291–2294).
- Broeder, D., & Wittenburg, P. (2006). The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies*, 1(2), 119–132.
- Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., . . . Wellner, P. (2006). The AMI meeting corpus: A pre-announcement. In *Proceedings of the second international conference on machine learning for multimodal interaction* (pp. 28–39). Berlin, Heidelberg: Springer-Verlag. Retrieved from http://dx.doi.org/10.1007/11677482_3 doi: 10.1007/11677482_3
- Carnap, R. (1972). A metafizika kiküszöbölése a nyelv logikai elemzésén keresztül. In F. Alt-richter (Ed.), *A bécsi kör filozófiája* (pp. 61–92). Gondolat.
- Clarke, D. D., & Argyle, M. (1997). Beszélgetési szekvenciák. In C. Pléh, I. Síklaki, & T. Terestyéni (Eds.), *Nyelv, kommunikáció, cselekvés* (pp. 565–602). Budapest: Osiris Kiadó.
- Couper-Kuhlen, E. (1991). A rhythm-based metric for turn-taking. In *12th International congress of phonetic science* (Vol. 1, p. 275-278). Aix-en-Provence: Université de Provence, Service de Publications.
- Couper-Kuhlen, E., & Selting, M. (1996). Towards an interactional perspective on prosody and a prosodic perspective on interaction. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: Interactional studies* (p. 11–56). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511597862.003

- Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.
- d'Alessandro, C., & Mertens, P. (1994). Automatic pitch contour stylization using a model of tonal perception. *Computer, Speech and Language*, 9(3), 257–288.
- de Jong, N. H., & Wempe, T. (2009, May 01). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2), 385–390. Retrieved from <https://doi.org/10.3758/BRM.41.2.385>
- De Looze, C., & Hirst. (2010). Integrating changes of register into automatic intonation analysis. In *Speech prosody 2010 conference*. Chicago.
- Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. In (Vol. 23, pp. 283–292.).
- Edlund, J., & Heldner, M. (2007). Underpinning /nailon/: Automatic estimation of pitch range and speaker relative pitch. In C. Müller (Ed.), *Speaker classification II: Selected projects* (pp. 229–242). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-540-74122-0_18
- Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.
- Fechner, G. T. (1860). *Elemente der psychophysik [Elements of psychophysics]*. Leipzig: Breitkopf und Härtel.
- Frid, J., Ambrazaitis, G., Schötcz, S., & Ambrazaitis, G. (2010). Automatic estimation of pitch range through distribution fitting. In *Proceedings from fonetics working papers 54.* (p. 41-46).
- Galibert, O. (2013, 08). Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech..
- Graf, S., Herbig, T., Buck, M., & Schmidt, G. (2015, 11). Features for voice activity detection: a comparative analysis. *EURASIP Journal on Advances in Signal Processing*, 2015, 91. doi: 10.1186/s13634-015-0277-z
- Gósy, M. (2000). A beszédritmus elemzésének egy lehetséges megközelítése. *Magyar Nyelvőr*, 124, 273–287.
- Gósy, M. (2004). *Fonetika, a beszéd tudománya*. Budapest: Osiris Kiadó.

- Gósy, M., Gyarmathy, D., Horváth, V., Grácz, T. E., Beke, A., Neuberger, T., & Nikléczy, P. (2012). BEA: Beszélt nyelvi adatbázis. In M. Gósy (Ed.), *Beszéd, adatbázis, kutatások* (pp. 9–24). Budapest: Akadémiai Kiadó.
- Hamill, T., & Price, L. L. (2008). *The hearing sciences*. San Diego: Plural Publishing.
- Hart, J. t. (1976). Psychoacoustic backgrounds of pitch contour stylisation. *IPO-APR*, 11, 11-19.
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298, 1569–1579.
- Hirschberg, J. (2002). Communication and prosody: functional aspects of prosody. *Speech Communication*, 36, 31–43.
- Hirschberg, J., & Nakatani, C. H. (1996). A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the ACL* (pp. 286–293).
- Hirschberg, J., & Nakatani, C. H. (1998). Acoustic indicators of topic segmentation. In *Proceedings of ICSLP*.
- Honbolygó, F. (2011). *A beszéd prozódiai jellemzőinek észlelése. a hangsúly pszicholingvisztikai és agyi háttere*. Budapest: Akadémia Kiadó.
- Huijbregts, M. (2008). *Segmentation, diarization and speech transcription: surprise data unraveled* (Unpublished doctoral dissertation). University of Twente, Netherlands.
- Hunyadi, L. (2002). *Hungarian sentence prosody and universal grammar*. Frankfurt: Peter Lang Verlag.
- Hunyadi, L. (2003a). A hangsúlyról a fonológiában és a fonetikában. In L. Hunyadi (Ed.), *Kísérleti fonetika, laboratóriumi fonológia a gyakorlatban* (pp. 65–91). Debrecen: Kossuth Egyetemi Kiadó.
- Hunyadi, L. (2003b). A logikai forma és a prozódia viszonya és e viszony univerzális implikációi. *Magyar nyelvjárások XLI*, 217–229.
- Hunyadi, L. (2011). Multimodal human-computer interaction technologies. theoretical modeling and application in speech processing. *Argumentum*, 7, 240–260.

- Hunyadi, L., Földesi, A., Szekrényes, I., Staudt, A., Kiss, H., Abuczki, A., & Bódog, A. (2012). Az ember–gép kommunikáció elméleti–technológiai modellje és nyelvtechnológiai vonatkozásai. In *Általános nyelvészeti tanulmányok XXIV: Nyelvtechnológiai kutatások* (pp. 265–309). Budapest: Akadémiai Kiadó.
- Hunyadi, L., Kiss, H., & Szekrényes, I. (2016a, 01). Incompleteness and fragmentation: Possible formal cues to cognitive processes behind spoken utterances. In (p. 231-257). doi: 10.1007/978-3-319-21209-8_14
- Hunyadi, L., Kiss, H., & Szekrényes, I. (2016b). Prosody enhances cognitive infocommunication: Materials from the hucomtech corpus. In A. Esposito & C. L. Jain (Eds.), *Toward robotic socially believable behaving systems - volume I: Modeling emotions* (pp. 183–204). Cham: Springer International Publishing.
- Hunyadi, L., Szekrényes, I., Czap, L., & Sziklai, I. (2014). Seeing the sounds? *Argumentum*, 10, 325–338.
- Hunyadi, L., Szekrényes, I., & Sziklai, I. (2015). Vizuális percepció és nyelvi feldolgozás. *Beszédkutatás*, 23, 186-208.
- Hámori, A., & Horváth, V. (2019). Társalgás, beszélőváltás és diskurzusszerveződés új megközelítésben – fonetikai jellemzők és pragmatikai tényezők összefüggései magyar társalgásokban (pilot study). *Beszédkutatás*, 27, 134-153.
- Ibrahim, N. S., & Ramli, D. A. (2018). I-vector extraction for speaker recognition based on dimensionality reduction. *Procedia Computer Science*, 126, 1534 - 1540. Retrieved from <http://www.sciencedirect.com/science/article/pii/S1877050918314042> doi: <https://doi.org/10.1016/j.procs.2018.08.126>
- Iványi, Z. (2001). A nyelvészeti konverzációelemzés. *Magyar Nyelvőr*, 125(1), 74-93.
- Jackendoff, R. (2002). *Foundations of language: brain, meaning, grammar, evolution*. Oxford: Oxford University Press.
- Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., & Mason, M. (2011). i-vector based speaker recognition on short utterances. In *Interspeech*.
- Kanrar, S. (2017). Speaker identification by GMM based i Vector. *CoRR*, abs/1704.03939. Retrieved from <http://arxiv.org/abs/1704.03939>

- Kassai, I. (1994). A fonetikai háttér. In *Struktúrális magyar nyelvtan 2. Fonológia* (pp. 581–696). Budapest: Akadémia Kiadó.
- Kiefer, F. (2009). Chomsky és a nyelvi jelentés kutatása. *Magyar Tudomány*, 170(9), 1037–1039.
- Kornai, A., & Kálmán, L. (1988). Hungarian intonation. In H. Hulst van Der & N. Smith (Eds.), *Autosegmental studies on pitch accent* (pp. 183–195). Dordrecht: Foris Publications.
- Kornai, A., & Szekrényes, I. (2017). e-Magyar beszédarchívum. In A. Tanács & V. Vincze (Eds.), *XIII. magyar számítógépes nyelvészeti konferencia (MSZNY2017)* (pp. 103–109). Szeged: JATEPress.
- Kovács, G., & Szekrényes, I. (2019). Applying neural network techniques for topic change detection in the hucomtech corpus. In L. Hunyadi & I. Szekrényes (Eds.), *The temporal structure of multimodal communication: Theory, methods and applications* (pp. 147–162). Cham: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-030-22895-8_8 doi: 10.1007/978-3-030-22895-8_8
- Kovács, G., & Váradi, T. (2017). A különböző modalitások hozzájárulásának vizsgálata a témairányítás eseteinek osztályozásához a hucomtech korpuszon. In A. Tanács & V. Vincze (Eds.), *XIII. magyar számítógépes nyelvészeti konferencia (MSZNY2017)* (pp. 103–109). Szeged: JATEPress.
- Kozima, H. (1993). Text segmentation based on similarity between words. In *Proceedings of ACL* (pp. 286–288).
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation. A bradford book*. Cambridge (Massachusetts)–London (England): The MIT Press.
- Luz, S., & Su, J. (2010). Assessing the effectiveness of conversational features for dialogue segmentation in medical team meetings and in the ami corpus. In *Proceedings of SIGDIAL* (p. 332-339).
- Magnusson, M. S. (2000, Mar 01). Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, & Computers*, 32(1), 93–110. Retrieved from <https://doi.org/10.3758/BF03200792>
- Markó, A. (2005). *A spontán beszéd néhány szupraszegmentális jellegzetessége*. (PhD dissertation). Eötvös Lóránd Tudományegyetem.

- Markó, A., & Gósy, M. (2015). A megszólalás stratégiái társalgásban. In V. Bárdosi (Ed.), *A nyelvi pragmatika kérdései szinkrón és diakrón megközelítésben* (pp. 159–168). Budapest: Tinta Könyvkiadó.
- Mertens, P. (2004). The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In *Proceedings of speech prosody*.
- Molugu, M. C. (2003). *Topic segmentation* (Unpublished master's thesis). University of Edinburgh.
- Nakajima, S., & Allen, J. F. (1997). Prosodic features of utterances in task-oriented dialogues. In *Computing PROSODY* (pp. 377–386). New York: Springer-Verlag.
- Nooteboom, S. (1997). The prosody of speech: Melody and rhythm. In *The handbook of phonetic sciences, nr. 5 in blackwell handbooks in linguistics, chap* (pp. 640–673).
- Németh, T. E. (2006). Az „azonos” nézetek különbözősége. *Magyar Nyelv*, 102, 420–429.
- Passonneau, R. J., & Litman, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1), 103–139.
- Pierrehumbert, J. B., & Hirschberg, J. (1990). The meaning of intonation contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Plans and intentions in communication and discourse* (pp. 271–311). MIT Press.
- Ping, A. (2005, 12). Talking themes: The thematic structure of talk. *Discourse Studies*, 7, 701–732. doi: 10.1177/1461445605055423
- Plug, L. (2009). On tempo in dispreferred turns: A recurrent pattern in a dutch corpus. In D. Barth-weingarten, N. Dehé, & A. Wichmann (Eds.), *Where prosody meets pragmatics* (p. 225–256). Bingley: Emerald.
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Vesely, K. (2011, December). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society. (IEEE Catalog No.: CFP11SRW-USB)
- Pápay, K., Szeghalmy, S., & Szekrényes, I. (2011). Hucomtech Multimodal Corpus annotation. *Arguemntum*, 7, 330–347.

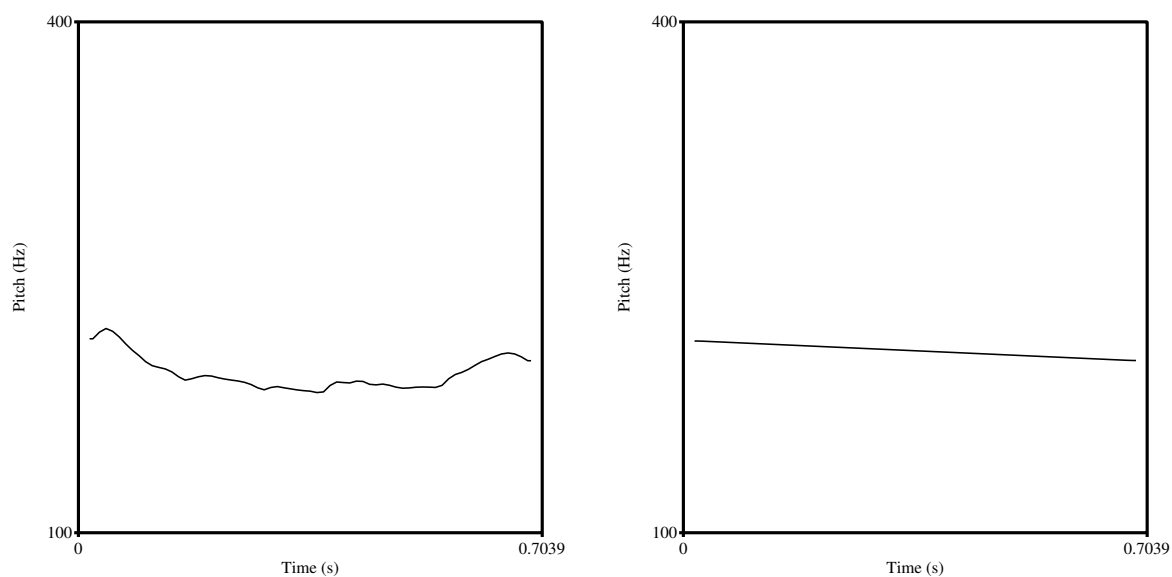
- Reichel, D., Uwe, Markó, A., & Mády, K. (2014). Parameterization and automatic labeling of hungarian intonation. In N. Campbell, D. Gibbon, & D. Hirst (Eds.), *7th International conference on speech prosody, SP 2014*. (pp. 738–742). Dublin: International Speech Communication Association (ISCA).
- Reynar, J. C. (1994). An automatic method of finding topic boundaries. In *Proceedings of ACL* (pp. 331–333).
- Roach, P. (2009). *English phonetics and phonology*. Cambridge: Cambridge University Press.
- Rosenberg, A. (2010). AuToBI – a tool for automatic ToBi annotation. In T. Kobayashi, K. Hirose, & S. Nakamura (Eds.), *Interspeech 2010* (pp. 146–149).
- Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., & Meignier, S. (2013, 08). An open-source state-of-the-art toolbox for broadcast news diarization..
- Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4, Part 1), 696–735. Retrieved from <http://www.jstor.org/stable/412243>
- Sahidullah, M., & Saha, G. (2012, 10). Comparison of speech activity detection techniques for speaker recognition. *CoRR*, *abs/1210.0297*.
- Schmidt, T. (2016). Good practices in the compilation of folk, the research and teaching corpus of spoken german. In J. M. Kirk & G. Andersen (Eds.), *Compilation, transcription, markup and annotation of spoken corpora, special issue of the international journal of corpus linguistics [IJCL 21:3]* (pp. 396–418).
- Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.
- Shriberg, E., Stolcke, A., Hakkani-Tür, D., & Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32, 127–154.
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., W. Wightman, C., Price, P., ... Hirschberg, J. (1992). ToBi: A standard for labeling English prosody. In *Proceedings of ICSLP* (pp. 867–870).
- Sitbon, L., & Bellot, P. (2007). Topic segmentation using weighted lexical links (WLL). In *Proceedings of the 30th Annual international ACM SIGIR conference on research and development in information retrieval* (pp. 737–738).

- Snyder, D., Chen, G., & Povey, D. (2015). *MUSAN: A music, speech, and noise corpus*. (arXiv:1510.08484v1)
- Snyder, D., Garcia-Romero, D., Sell, G., McCree, A., Povey, D., & Khudanpur, S. (2019, May). Speaker recognition for multi-speaker conversations using X-vectors. In *ICASSP 2019 - 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (p. 5796-5800).
- Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors: Robust DNN embeddings for speaker recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329-5333.
- Szaszák, G. (2010). A prozódia szerepe a beszédfelismerésben. In *A magyar beszéd: Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek* (pp. 390–392). Budapest: Akadémiai Kiadó.
- Szekrényes, I., & Kovács, G. (2017). Classification of formal and informal dialogues based on turn-taking and intonation using deep neural networks. In A. Karpov, R. Potapova, & I. Mporas (Eds.), *Speech and computer* (pp. 233–243). Cham: Springer International Publishing.
- Szekrényes, I. (2011). Automatizált f0-mérések beágyazott mondat szerkezetek rekurziós sajátosságainak vizsgálatához. In L. Hunyadi (Ed.), *Rekurzió a nyelvben I.: Prozódiai megközelítés* (pp. 141–152). Budapest.
- Szekrényes, I. (2014). Annotation and interpretation of prosodic data in the hucomtech corpus for multimodal user interfaces. *Journal on Multimodal User Interfaces*, 8:(2), 143–150.
- Szekrényes, I. (2015). Prosotool, a method for automatic annotation of fundamental frequency. In *6th IEEE International conference on cognitive Infocommunications (CogInfoCom)* (pp. 291–296). New York: IEEE.
- Szekrényes, I. (2019). Post-processing t-patterns using external tools from a mixed method perspective. *Frontiers in Psychology*, 10, 1-11. Retrieved from <https://www.frontiersin.org/article/10.3389/fpsyg.2019.01680> doi: 10.3389/fpsyg.2019.01680
- Szekrényes, I., Csipkés, L., & Oravecz, C. (2011). A HuComTech-korpusz és -adatbázis számítógépes feldolgozási lehetőségei, automatikus prozódiai annotáció. In V. Vincze (Ed.), *VIII. magyar számítógépes nyelvészeti konferencia MSZNY2011* (pp. 190–198). Szeged: JATEPress.

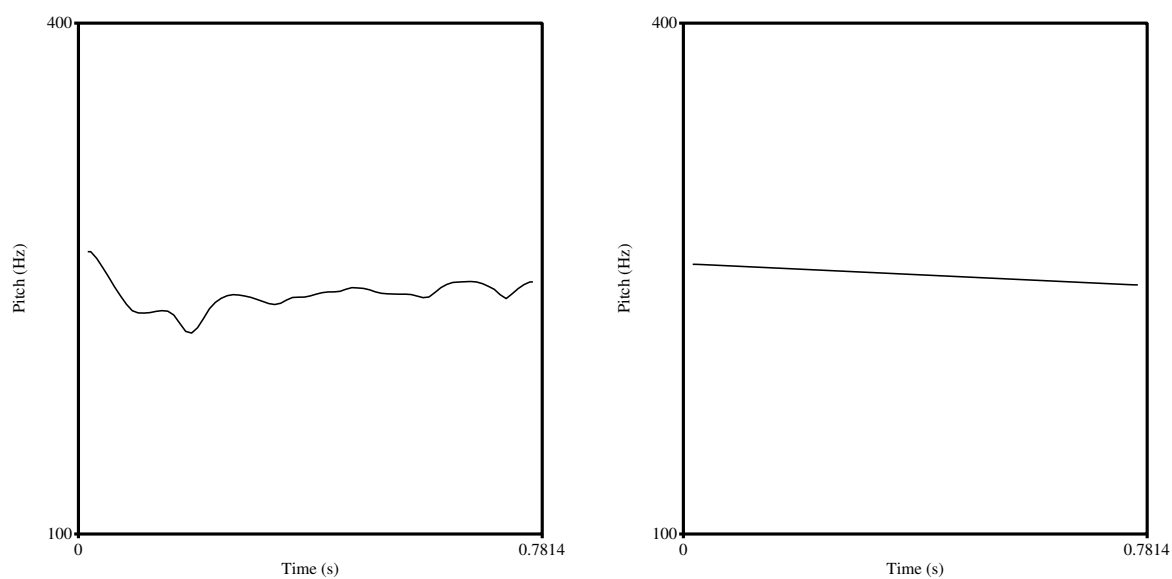
- Tarnóczy, T. (1984). *Hangnyomás, hangosság, zajosság*. Budapest: Akadémiai Kiadó.
- Taylor, P. (2000). Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107(3), 1697–1714.
- Varga, L. (1994). A hanglejtés. In *Struktúrális magyar nyelvtan 2. Fonológia* (pp. 468–549). Budapest: Akadémia Kiadó.
- Vozáriková, E., & Juhár, J. (2015, 11). Comparison of diarization tools for building speaker database. *Advances in Electrical and Electronic Engineering*, 13, 314–319. doi: 10.15598/aeec.v13i4.1468
- Váradi, T., Simon, E., Sass, B., Gerőcs, M., Mittelholcz, I., Novák, A., ... Vincze, V. (2017). Az e-magyar digitális nyelvfeldolgozó rendszer. In V. Vincze (Ed.), *XIII. magyar számítógépes nyelvészeti konferencia (mszny2017)* (pp. 49–60). Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport.
- Withers, P. (2012). Metadata management with Arbil. In V. Arranz, D. Broeder, B. Gaiffe, M. Gavrilidou, & M. Monachini (Eds.), *Proceedings of the workshop describing LRs with metadata: Towards flexibility and interoperability in the documentation of LR at LREC 2012* (pp. 72–75). ELRA.
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of LREC 2006* (pp. 213–269).
- Zajíc, Z., Kunesová, M., & Radová, V. (2016). Investigation of segmentation in i-Vector based speaker diarization of telephone speech. In *Specom*.
- Zellers, M. (2012). Prosodic variation for topic shift and other functions in local contrasts in conversation. *Phonetica*, 69 4, 231–53.
- Zellers, M., & Post, B. (2009). Fundamental frequency and other prosodic cues to topic structure. In *Proceedings of IDP* (pp. 377–386). Paris.

Melléklet

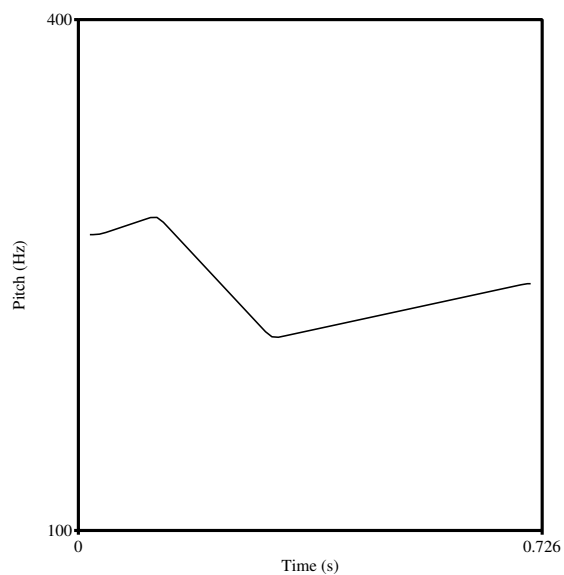
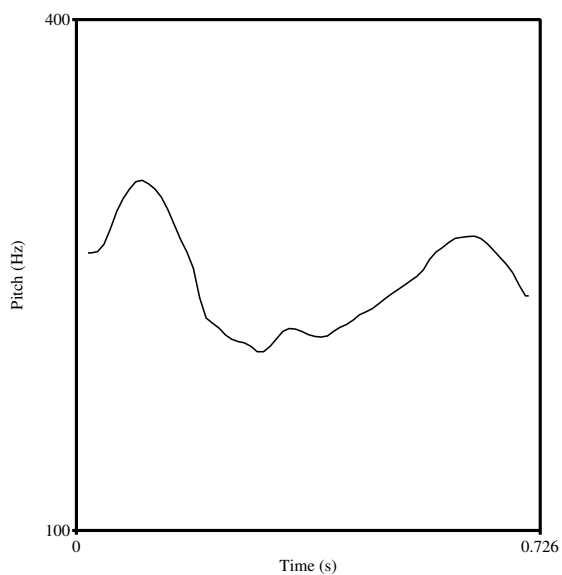
Az 5.3. fejezethez tartozó ábrák



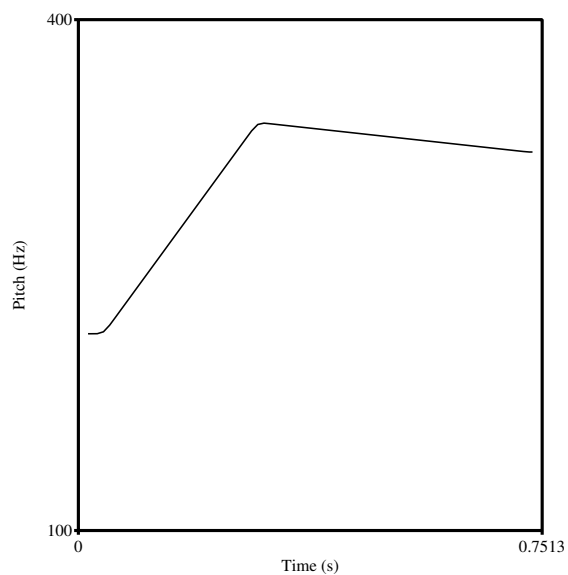
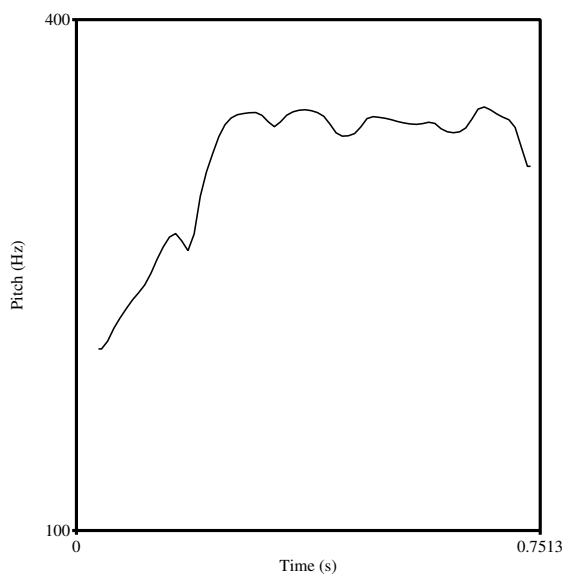
1. ábra. Az „Angéla néni” megnyilatkozás 1. variációja (önálló + befejezett) alapján kapott F0 görbe (balra) és annak stilizált verziója (jobbra)



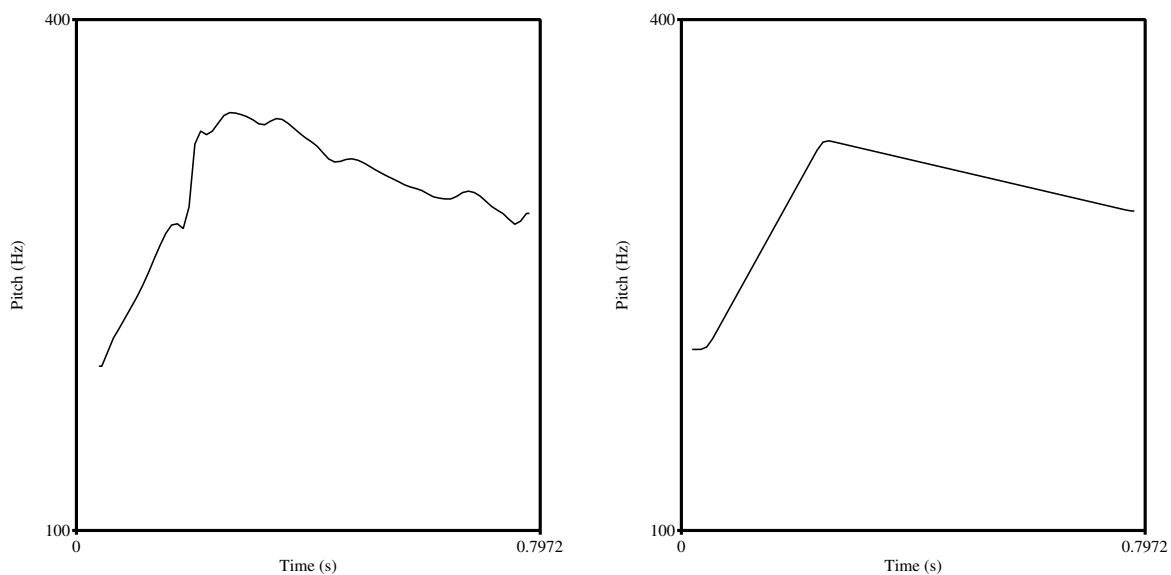
2. ábra. Az „Angéla néni” megnyilatkozás 2. variációja (önálló) alapján kapott F0 görbe (balra) és annak stilizált verziója (jobbra)



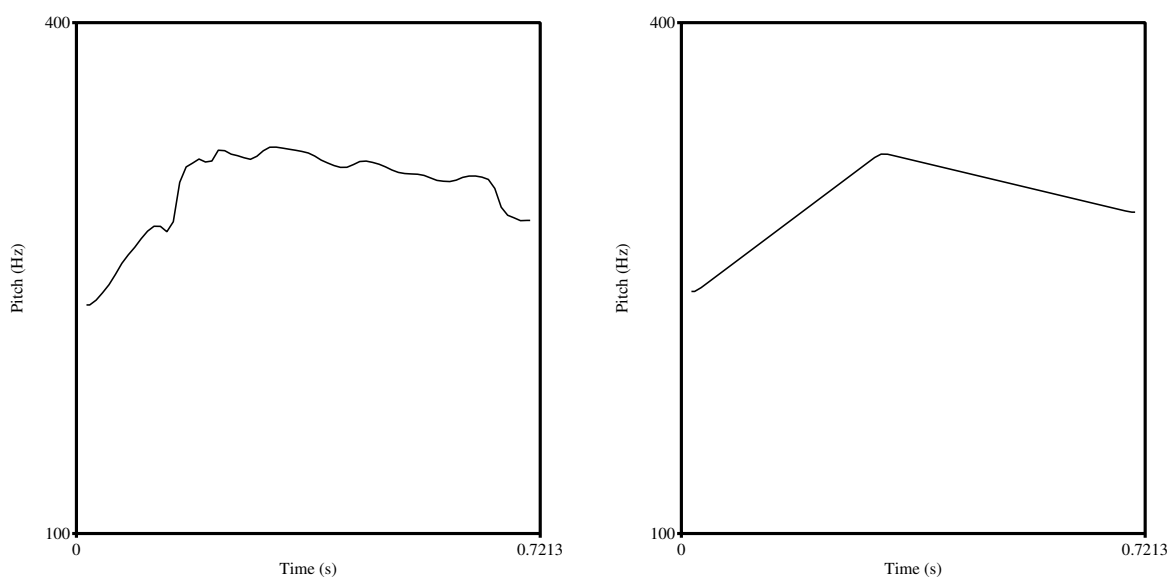
3. ábra. Az „Angéla néni” megnyilatkozás 3. variációja (önálló + szembehelyezkedő) alapján kapott F0 görbe (balra) és annak stilizált verziója (jobbra)



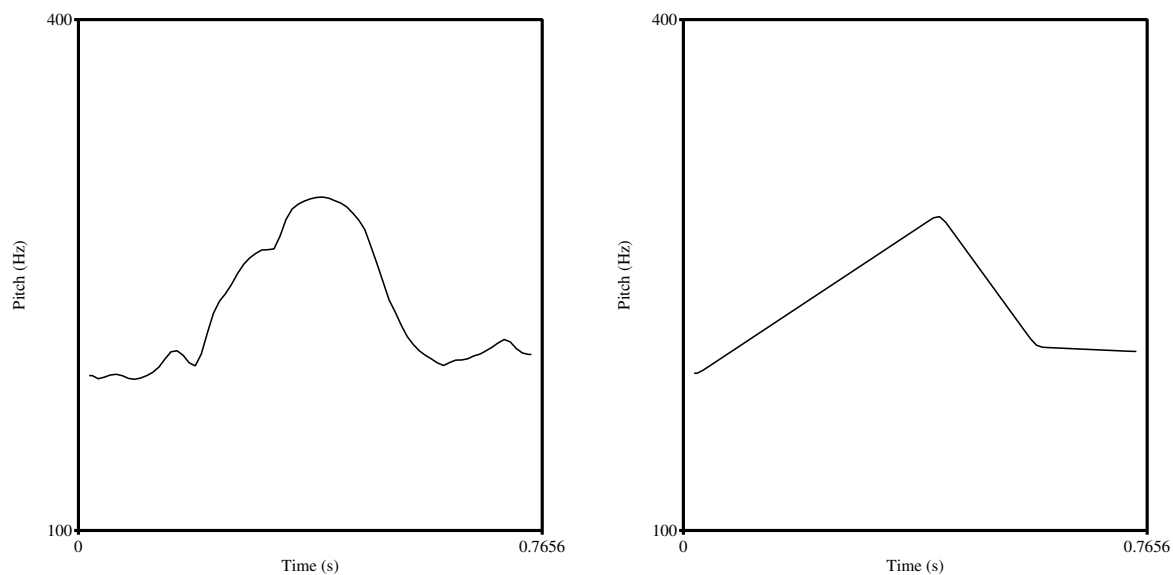
4. ábra. Az „Angéla néni” megnyilatkozás 4. variációja (előre mutató) alapján kapott F0 görbe (balra) és annak stilizált verziója (jobbra)



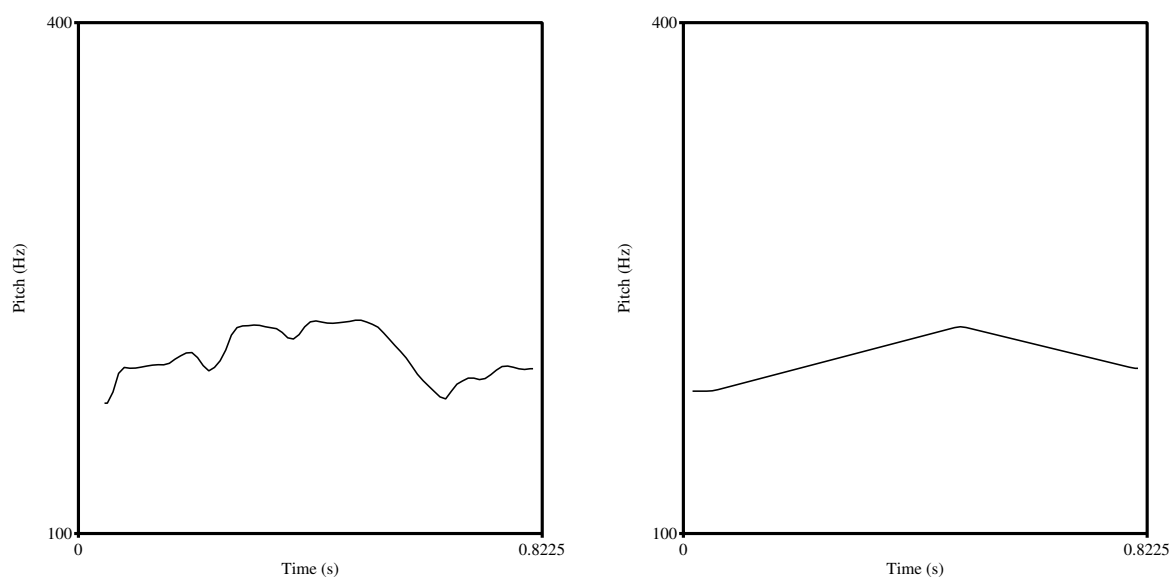
5. ábra. Az „Angéla néni” megnyilatkozás 5. variációja (előre mutató + feszült) alapján kapott F0 görbe (balra) és annak stilizált verziója (jobbra)



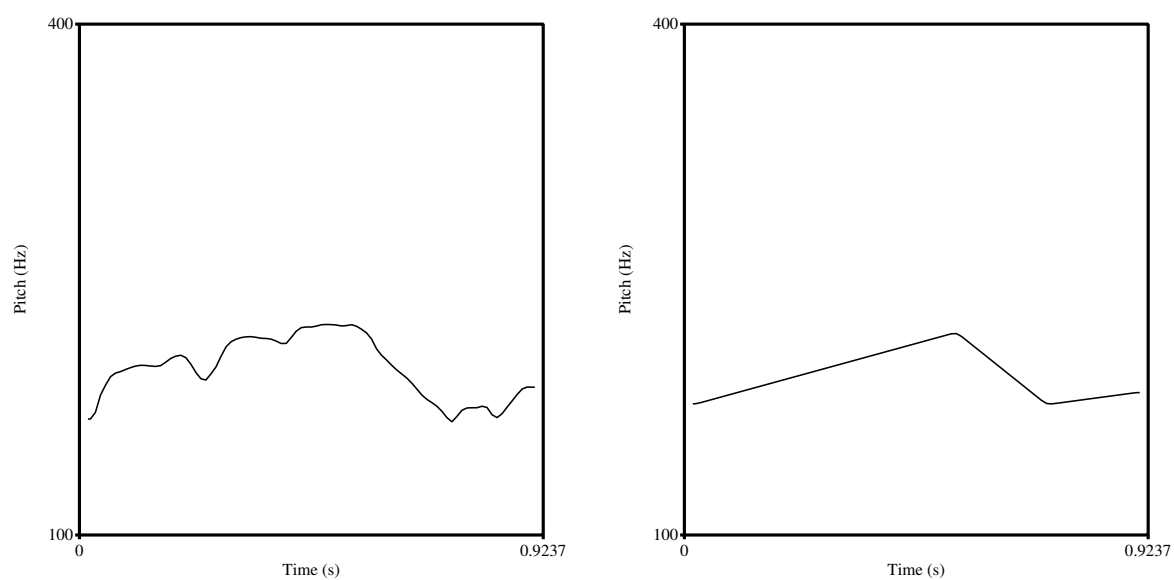
6. ábra. Az „Angéla néni” megnyilatkozás 6. variációja (előre mutató + rutinszerű) alapján kapott F0 görbe (balra) és annak stilizált verziója (jobbra)



7. ábra. Az „Angéla néni” megnyilatkozás 7. variációja (eldöntendő kérdő + kérdező) alapján kapott F0 görbe (balra) és annak stilizált verziója (jobbra)



8. ábra. Az „Angéla néni” megnyilatkozás 8. variációja (eldöntendő kérdő) alapján kapott F0 görbe (balra) és annak stilizált verziója (jobbra)



9. ábra. Az „Angéla néni” megnyilatkozás 9. variációja (eldöntendő kérdő + felkiáltó) alapján kapott F0 görbe (balra) és annak stilizált verziója (jobbra)

Az értekezés magyar nyelvű összefoglalója

A disszertációm célja, hogy a beszélt nyelv kutatásához olyan újszerű módszertani eszközökkel és kísérletekkel járuljon hozzá, amelyek a beszéd prozódiaját nem csak a megnyilatkozások szintjén, hanem a társas interakciók mikro- és makrostruktúrája felől, mint önmagában is informatív aktusok mintázatát közelítik meg.

Az értekezés elején tágabb perspektívából elemzem, hogy a nyelvtudomány különféle irányzatai (generatív grammatika, beszédaktus-elmélet, konverzációs analízis) a prozódia milyen aspektusait világítják meg, majd bemutatom azokat a konkrét elméleti kérdéseket és hipotéziseket, amelyek a disszertáció keretében megvalósított kutatást leginkább motiválták. A saját fejlesztéseim és kísérleteim ismertetése előtt röviden körbejárom a beszédtechnológia kapcsolódó területeit (beszéddetektálás és beszélő diarizálás) is, illetve ismertetem az intonáció modellezéséhez és számítógépes feldolgozásához elérhető megoldásokat és keretrendszerket (ToBI, Tilt, Prosogram).

A disszertáció legrészletesebben a prozódia automatikus annotációjára kidolgozott módszerrel foglalkozik, amelynek célja, hogy a beszéd dallamának, intenzitásának és tempójának modulációit egymást követő, kategorikus címkékkel jellemzett események szekvenciájaként írja le. Technikai szempontból ez az adott jellemzőt leginkább meghatározó fizikai paraméter (a dallama esetén például a beszédjel alaphfrekvenciája) mérhető értékeinek stilizálását, illetve a stilizáció eredményeként kapott, a moduláció fő irányát kifejező trendek elemzését követelte meg. Hipotézisem szerint a kimenetként kapott esemény-szekvenciák olyan, a percepció számára is releváns modulációkat tárnak fel, amelyek egy megnyilatkozás attitűdtartalmának értelmezéséhez, a társalgás tagolásához, vagy hosszabb mintákban a beszédszituáció jellemzéséhez is támpontokat jelenthetnek. A beszéd dallamának elemzését végző algoritmus szöveges kimenetét webes környezetben megjeleníthető vizualizációkkal is kiegészítettem. Ez részben az algoritmus validálása érdekében történt, hogy a kimenetként kapott dallam-kontúrok könnyebben összevethetők legyenek a percepcióban megjelenő hangélménnyel.

A hipotézisek igazolására több percepciós kísérletet is lebonyolítottam. Ezekben azt a kérdést vizsgáltam, hogy kizárólag a prozódiaira támaszkodva az alanyok (1) egy zárt halmazon belül meg tudják-e határozni a társalgási szituációt, (2) érzékelik-e a téma megváltozását, (3) tudják-e értékelni a az interakcióban résztvevők viszonyát, (4) tudják-e a társalgáson belül pozicionálni annak egy részletét, (5) be tudják-e azonosítani egy megnyilatkozás attitűdtartalmát. A témaváltással detektálásával és a beszédszituáció klasszifikálásával kapcsolatos feladatot a HuComTech korpuszból kinyert prozódiai jellemzőkön betanított mély neuronhálók (DNN) is kipróbáltam.

Az értekezés angol nyelvű összefoglalója

The aim of the dissertation is to contribute to the research of spoken language with new methodological tools and experiments that represent an approach to the prosody of speech not only at the level of utterances but also at that of the micro- and macrostructure of social interaction, as patterns of acts informative by themselves.

At the beginning of the dissertation, from a wider perspective, I analyze what aspects of prosody are treated by various schools of linguistics (generative grammar, speech act theory, conversation analysis), to be followed by specifying those concrete questions and hypotheses which motivated my research carried out within the frameworks of the dissertation. Before describing my own developments and experiments I give a short overview of related fields of language technology (speech detection and speaker diarization), and a summary of solutions and frameworks available for the modeling and computational processing of intonation (ToBI, Tilt, Prosogram).

The most detailed discussion in the dissertation is related to my methodology developed for the automatic annotation of prosody, whose aim it is to describe the modulations of speech melody, intensity and tempo as sequences of adjacent events characterized by categorical labels. For a technical point of view, it consisted of the stylization of the measurable values of the physical parameters of the given feature (as in the case of tone: the fundamental frequency of the speech signal), as well as the analysis of the trends resulting from the stylization, expressing the main direction of the modulation. According to my hypothesis, the resulting sequences of events discover those modulations relevant to perception which can serve as indicators for the interpretation of the attitudinal content of an utterance, the segmentation of a conversation, or, in longer samples, the characterization of the context of a given interaction. In addition, I implemented a web based visualization of the textual output of the analysis of speech melody. It was partly done for the purpose of validation so that the output tonal contours could be more easily comparable with the perceived sound experience.

I carried out a number of experiments for the validation of the hypotheses. Through them I wished to find an answer to the questions whether the subjects (1) can determine the situation of the conversation, (2) sense if the topic has changed, (3) can give a judgement about the relation between participants of the interaction, (4) can position within the conversation one of its given segments, (5) can identify the attitudinal content of an utterance. I also tested the task of the classification of the conversation situation training deep neuronal networks (DNNs) based on prosodic features included in the HuComTech corpus.



Nyilvántartási szám: DEENK/367/2019.PL
Tárgy: PhD Publikációs Lista

Jelölt: Szekrényes István
Neptun kód: SQGGTJ
Doktori Iskola: Nyelvtudományok Doktori Iskola
MTMT azonosító: 10038180

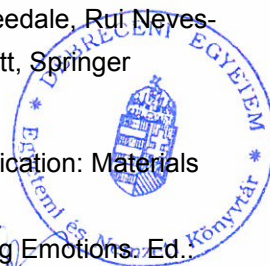
A PhD értekezés alapjául szolgáló közlemények

Magyar nyelvű könyvrészletek (2)

1. Hunyadi, L., Földesi, A., **Szekrényes, I.**, Staudt, A., Kiss, H., Abuczki, Á., Bódog, A.: Az ember-gép kommunikáció elméleti-technológiai modellje és nyelvtechnológiai vonatkozásai.
In: Általános Nyelvészeti Tanulmányok XXIV. Nyelvtechnológiai kutatások. Szerk.: Kenesei I., Prószték G., Várady T, Akadémiai Kiadó, Budapest, 265-309, 2012. ISBN: 9789630593083
2. **Szekrényes, I.**: Automatizált F0-mérések beágyazott mondatstruktúrák rekurziós sajátosságainak vizsgálatához.
In: Rekúzió a nyelvben I. Prozódiák megközelítés. Szerk.: Hunyadi László, Tinta Könyvkiadó, Budapest, 156-169, 2011, (Segédkönyvek a nyelvészet tanulmányozásához, ISSN 1419-6603 ; 123.) ISBN: 9789639902824

Idegen nyelvű, külföldi könyvrészletek (3)

3. Kovács, G., **Szekrényes, I.**: Applying Neural Network Techniques for Topic Change Detection in the HuComTech Corpus.
In: The Temporal Structure of Multimodal Communication: Theory, Methods and Applications. Ed.: Hunyadi László, Szekrényes István, Springer International Publishing Ag, Cham, 147-162, 2019. ISBN: 9783030228941
4. Hunyadi, L., Kiss, H., **Szekrényes, I.**: Incompleteness and Fragmentation: Possible Formal Cues to Cognitive Processes Behind Spoken Utterances.
In: Intelligent Decision Technology Support in Practice. Ed.: Jeffrey W. Tweedale, Rui Neves-Silva, Lakhmi C. Jain, Gloria Phillips-Wren, Junzo Watada, Robert J Howlett, Springer International Publishing Ag, Cham, 231-257, 2016. ISBN: 9783319212081
5. Hunyadi, L., **Szekrényes, I.**, Kiss, H.: Prosody Enhances Cognitive Infocommunication: Materials from the HuComTech Corpus.
In: Toward Robotic Socially Believable Behaving Systems. Vol. I : Modeling Emotions. Ed.: Anna Esposito, Lakhmi C. Jain, Springer International Publishing Ag, Cham, 183-204, 2016. ISBN: 9783319310565





Magyar nyelvű tudományos közlemények hazai folyóiratban (1)

6. Hunyadi, L., **Szekrényes, I.**, Sziklai, I.: Vizuális percepció és nyelvi feldolgozás.

Beszéd kutatás. 23, 186-208, 2015. ISSN: 1218-8727.

Idegen nyelvű tudományos közlemények hazai folyóiratban (2)

7. Hunyadi, L., **Szekrényes, I.**, Czap, L., Sziklai, I.: Seeing the sounds?

Argumentum (Debr.). 10, 325-338, 2014. EISSN: 1787-3606.

8. Pápay, K., Szeghalmy, S., **Szekrényes, I.**: HuComTech Multimodal Corpus Annotation.

Argumentum (Debr.). 7, 330-347, 2011. EISSN: 1787-3606.

Idegen nyelvű tudományos közlemények külföldi folyóiratban (2)

9. **Szekrényes, I.**: Post-processing T-patterns Using External Tools From a Mixed Method Perspective.

Front. Psychol. 10, 1-12, 2019. EISSN: 1664-1078.

DOI: <http://dx.doi.org/10.3389/fpsyg.2019.01680>

IF: 2.129 (2018)

10. **Szekrényes, I.**: Annotation and interpretation of prosodic data in the HuComTech corpus for multimodal user interfaces.

J. Multimodal User Interfaces. 8 (2), 143-150, 2014. ISSN: 1783-7677.

DOI: <http://dx.doi.org/10.1007/s12193-013-0140-1>

IF: 0.797

Magyar nyelvű konferencia közlemények (2)

11. Kornai, A., **Szekrényes, I.**: e-Magyar beszédarchívum.

In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017). Szerk.: Vincze Veronika, Szegedi Tudományegyetem, Szeged, 103-109, 2017. ISBN: 9789633065181

12. **Szekrényes, I.**, Csipkés, L., Oravecz, C.: A HuComTech-korpusz és -adatbázis számítógépes feldolgozási lehetőségei. Automatikus prozódiai annotáció.

In: VIII. Magyar Számítógépes Nyelvészeti Konferencia : MSZNY 2011. Szerk.: Tanács Attila, Vincze Veronika, Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, 190-198, 2011. ISBN: 9789633061213

Idegen nyelvű konferencia közlemények (6)

13. Hunyadi, L., Váradi, T., Kovács, G., **Szekrényes, I.**, Kiss, H., Takács, K.: Human-human, human-machine communication: on the HuComTech multimodal corpus.

In: Selected papers from the CLARIN Annual Conference 2018. Ed.: Inguna Skadina, Maria Eskevich, Linköping University Electronic Press, Linköping, 56-65, 2018, (Linköping Electronic Conference Proceedings, ISSN 1650-3686) ISBN: 9789176850343





14. **Szekrényes, I.**, Kovács, G.: Classification of Formal and Informal Dialogues Based on Turn-Taking and Intonation Using Deep Neural Networks.
In: Speech and Computer : 19th International Conference, SPECOM 2017: Proceedings. Ed.: Alexey Karpov, Rodmonga Potapova, Iosif Mporas, Springer International Publishing Ag, Cham, 233-243, 2017. ISBN: 9783319664286
15. Hunyadi, L., Váradi, T., **Szekrényes, I.**: Language technology tools and resources for the analysis of multimodal communication.
In: Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH). Ed.: Erhard Hinrichs, Marie Hinrichs, Thorsten Trippel, University of Tübingen, Tübingen, 117-124, 2016. ISBN: 9784879747082
16. **Szekrényes, I.**: ProsoTool, a method for automatic annotation of fundamental frequency.
In: 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). Ed.: Baranyi Péter, Csapó Ádám, IEEE, New York, 291-296, 2015.
17. Hunyadi, L., **Szekrényes, I.**, Borbély, A., Kiss, H.: Annotation of spoken syntax in relation to prosody and multimodal pragmatics.
In: 3rd IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2012 : Proceedings, December 2-5, 2012 Košice, Slovakia. Ed.: Péter Baranyi, IEEE, Danvers, 537-541, 2012. ISBN: 9781467351881
18. Esfandiari Baiat, G., **Szekrényes, I.**: Topic change detection based on prosodic cues in unimodal setting.
In: 3rd IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2012 : Proceedings, December 2-5, 2012 Košice, Slovakia. Ed.: Péter Baranyi, IEEE, Danvers, 527-530, 2012. ISBN: 9781467351874

További közlemények

Idegen nyelvű, hazai könyvek (1)

19. Csűry, I., Hunyadi, L., Abuczki, Á., Esfandiari Baiat, G., Földesi, A., **Szekrényes, I.**: Elements of Electronic Information and Document Processing: An introduction to informatics (not only) for the humanities. Debrecen Univ. Press, Debrecen, 139 p., 2016. ISBN: 9789633485643

Magyar nyelvű tudományos közlemények hazai folyóiratban (2)

20. **Szekrényes, I.**: Egy apolitikusan felfogott élet esélyei.
Nagyerdei Almanach. 1, 1-8, 2010. EISSN: 2062-3305.
21. **Szekrényes, I.**: A király meztelen: avagy Egy boldogtalan kritikus "apológiája".
Forrás. 39 (7-8), 161-167, 2007. ISSN: 0133-056X.





Idegen nyelvű konferencia közlemények (1)

22. Hunyadi, L., Bertók, K., Németh, T. E., **Szekrényes, I.**, Abuczki, Á., Nagy, G., Nagy, N., Németi, P., Bódog, A.: The outlines of a theory and technology of human-computer interaction as represented in the model of the HuComTech project.
In: Proceedings of 2nd Cognitive Infocommunications International Conference. Ed.: by Baranyi Péter, IEEE Computer Society, New York, 249-255, 2011. ISBN: 9781457718069

A közlő folyóiratok összesített impakt faktora: 2,926

**A közlő folyóiratok összesített impakt faktora (az értekezés alapjául szolgáló közleményekre):
2,926**

A DEENK a Jelölt által az iDEa Tudóstérbe feltöltött adatok bibliográfiai és tudománymetriai ellenőrzését a tudományos adatbázisok és a Journal Citation Reports Impact Factor lista alapján elvégezte.

Debrecen, 2019.11.06.





Registry number: DEENK/367/2019.PL
Subject: PhD Publikációs Lista

Candidate: István Szekrényes
Neptun ID: SQGGTJ
Doctoral School: Doctoral School of Linguistics
MTMT ID: 10038180

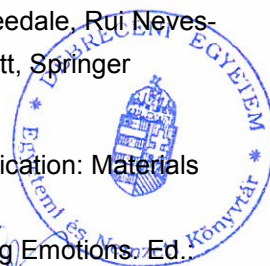
List of publications related to the dissertation

Hungarian book chapters (2)

1. Hunyadi, L., Földesi, A., **Szekrényes, I.**, Staudt, A., Kiss, H., Abuczki, Á., Bódog, A.: Az ember-gép kommunikáció elméleti-technológiai modellje és nyelvtechnológiai vonatkozásai.
In: Általános Nyelvészeti Tanulmányok XXIV. Nyelvtechnológiai kutatások. Szerk.: Kenesei I., Prószték G., Várady T, Akadémiai Kiadó, Budapest, 265-309, 2012. ISBN: 9789630593083
2. **Szekrényes, I.**: Automatizált F0-mérések beágyazott mondat szerkezetek rekurziós sajátosságainak vizsgálatához.
In: Rekurzió a nyelvben I. Prozódiák megközelítés. Szerk.: Hunyadi László, Tinta Könyvkiadó, Budapest, 156-169, 2011, (Segédkönyvek a nyelvészet tanulmányozásához, ISSN 1419-6603 ; 123.) ISBN: 9789639902824

Foreign language international book chapters (3)

3. Kovács, G., **Szekrényes, I.**: Applying Neural Network Techniques for Topic Change Detection in the HuComTech Corpus.
In: The Temporal Structure of Multimodal Communication: Theory, Methods and Applications. Ed.: Hunyadi László, Szekrényes István, Springer International Publishing Ag, Cham, 147-162, 2019. ISBN: 9783030228941
4. Hunyadi, L., Kiss, H., **Szekrényes, I.**: Incompleteness and Fragmentation: Possible Formal Cues to Cognitive Processes Behind Spoken Utterances.
In: Intelligent Decision Technology Support in Practice. Ed.: Jeffrey W. Tweedale, Rui Neves-Silva, Lakhmi C. Jain, Gloria Phillips-Wren, Junzo Watada, Robert J Howlett, Springer International Publishing Ag, Cham, 231-257, 2016. ISBN: 9783319212081
5. Hunyadi, L., **Szekrényes, I.**, Kiss, H.: Prosody Enhances Cognitive Infocommunication: Materials from the HuComTech Corpus.
In: Toward Robotic Socially Believable Behaving Systems. Vol. I : Modeling Emotions. Ed.: Anna Esposito, Lakhmi C. Jain, Springer International Publishing Ag, Cham, 183-204, 2016. ISBN: 9783319310565





Hungarian scientific articles in Hungarian journals (1)

6. Hunyadi, L., **Szekrényes, I.**, Sziklai, I.: Vizuális percepció és nyelvi feldolgozás.
Beszéd kutatás. 23, 186-208, 2015. ISSN: 1218-8727.

Foreign language scientific articles in Hungarian journals (2)

7. Hunyadi, L., **Szekrényes, I.**, Czap, L., Sziklai, I.: Seeing the sounds?
Argumentum (Debr.). 10, 325-338, 2014. EISSN: 1787-3606.
8. Pápay, K., Szeghalmy, S., **Szekrényes, I.**: HuComTech Multimodal Corpus Annotation.
Argumentum (Debr.). 7, 330-347, 2011. EISSN: 1787-3606.

Foreign language scientific articles in international journals (2)

9. **Szekrényes, I.**: Post-processing T-patterns Using External Tools From a Mixed Method Perspective.
Front. Psychol. 10, 1-12, 2019. EISSN: 1664-1078.
DOI: <http://dx.doi.org/10.3389/fpsyg.2019.01680>
IF: 2.129 (2018)
10. **Szekrényes, I.**: Annotation and interpretation of prosodic data in the HuComTech corpus for multimodal user interfaces.
J. Multimodal User Interfaces. 8 (2), 143-150, 2014. ISSN: 1783-7677.
DOI: <http://dx.doi.org/10.1007/s12193-013-0140-1>
IF: 0.797

Hungarian conference proceedings (2)

11. Kornai, A., **Szekrényes, I.**: e-Magyar beszédarchívum.
In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017). Szerk.: Vincze Veronika, Szegedi Tudományegyetem, Szeged, 103-109, 2017. ISBN: 9789633065181
12. **Szekrényes, I.**, Csipkés, L., Oravecz, C.: A HuComTech-korpusz és -adatbázis számítógépes feldolgozási lehetőségei. Automatikus prozódiai annotáció.
In: VIII. Magyar Számítógépes Nyelvészeti Konferencia : MSZNY 2011. Szerk.: Tanács Attila, Vincze Veronika, Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, 190-198, 2011. ISBN: 9789633061213

Foreign language conference proceedings (6)

13. Hunyadi, L., Váradi, T., Kovács, G., **Szekrényes, I.**, Kiss, H., Takács, K.: Human-human, human-machine communication: on the HuComTech multimodal corpus.
In: Selected papers from the CLARIN Annual Conference 2018. Ed.: Inguna Skadina, Maria Eskevich, Linköping University Electronic Press, Linköping, 56-65, 2018, (Linköping Electronic Conference Proceedings, ISSN 1650-3686) ISBN: 9789176850343





14. **Szekrényes, I.**, Kovács, G.: Classification of Formal and Informal Dialogues Based on Turn-Taking and Intonation Using Deep Neural Networks.
In: Speech and Computer : 19th International Conference, SPECOM 2017: Proceedings. Ed.: Alexey Karpov, Rodmonga Potapova, Iosif Mporas, Springer International Publishing Ag, Cham, 233-243, 2017. ISBN: 9783319664286
15. Hunyadi, L., Váradi, T., **Szekrényes, I.**: Language technology tools and resources for the analysis of multimodal communication.
In: Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH). Ed.: Erhard Hinrichs, Marie Hinrichs, Thorsten Trippel, University of Tübingen, Tübingen, 117-124, 2016. ISBN: 9784879747082
16. **Szekrényes, I.**: ProsoTool, a method for automatic annotation of fundamental frequency.
In: 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). Ed.: Baranyi Péter, Csapó Ádám, IEEE, New York, 291-296, 2015.
17. Hunyadi, L., **Szekrényes, I.**, Borbély, A., Kiss, H.: Annotation of spoken syntax in relation to prosody and multimodal pragmatics.
In: 3rd IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2012 : Proceedings, December 2-5, 2012 Košice, Slovakia. Ed.: Péter Baranyi, IEEE, Danvers, 537-541, 2012. ISBN: 9781467351881
18. Esfandiari Baiat, G., **Szekrényes, I.**: Topic change detection based on prosodic cues in unimodal setting.
In: 3rd IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2012 : Proceedings, December 2-5, 2012 Košice, Slovakia. Ed.: Péter Baranyi, IEEE, Danvers, 527-530, 2012. ISBN: 9781467351874

List of other publications

Foreign language Hungarian books (1)

19. Csúry, I., Hunyadi, L., Abuczki, Á., Esfandiari Baiat, G., Földesi, A., **Szekrényes, I.**: Elements of Electronic Information and Document Processing: An introduction to informatics (not only) for the humanities. Debrecen Univ. Press, Debrecen, 139 p., 2016. ISBN: 9789633185643

Hungarian scientific articles in Hungarian journals (2)

20. **Szekrényes, I.**: Egy apolitikusan felfogott élet esélyei.
Nagyerdei Almanach. 1, 1-8, 2010. EISSN: 2062-3305.
21. **Szekrényes, I.**: A király meztelen: avagy Egy boldogtalan kritikus "apológiája".
Forrás. 39 (7-8), 161-167, 2007. ISSN: 0133-056X.





Foreign language conference proceedings (1)

22. Hunyadi, L., Bertók, K., Németh, T. E., **Szekrényes, I.**, Abuczki, Á., Nagy, G., Nagy, N., Némethi, P., Bódog, A.: The outlines of a theory and technology of human-computer interaction as represented in the model of the HuComTech project.

In: Proceedings of 2nd Cognitive Infocommunications International Conference. Ed.: by Baranyi Péter, IEEE Computer Society, New York, 249-255, 2011. ISBN: 9781457718069

Total IF of journals (all publications): 2,926

Total IF of journals (publications related to the dissertation): 2,926

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

06 November, 2019

