Zsigri Gyula,* Tóth László,+ Kocsor András,+ Sejtes Györgyi*

Az automata és kézi szegmentálás ejtésvariációk okozta problémái

* Szegedi Tudományegyetem Magyar Nyelvészeti Tanszék
+MTA-SZTE Mesterséges Intelligencia Kutatócsoport

A beszédadatbázisok egyik legértékesebb része a beszédhang szintű szegmentálási információ. A szegmentálást és címkézést tökéletesen csakis nagy figyelmet igénylő, fáradságos és hosszadalmas kézi munkával lehet elvégezni. Megkönnyítheti és meggyorsíthatja viszont a munkát egy megfelelő, speciálisan erre a célra kialakított algoritmus, amely megkísérli automatikusan elhelyezni a fonetikai határokat. Akár ember, akár gép végzi a szegmentálást, segítségként rendelkezésére áll a hanganyag feltételezett fonetikai átirata, amelyet egy fonetikus átíró algoritmus állít elő a betű szerinti lejegyzésből. A jel valódi fonetikai tartalma azonban eltérhet ettől, hiszen ugyanannak a szövegnek ejtésvariációja lehet. A cikkben megvizsgáljuk, hogy ez a jelenség hogyan befolyásolja az általunk alkalmazott automata, illetve félautomata szegmentáló algoritmusokat. Megnézzük továbbá, hogy az MTBA adatbázis kézi feldolgozása során a szegmentálást végző személyek miben tértek el az előzetesen rögzített szabályoktól, különös tekintettel arra, hogy mentális (fonetikai) lexikonjuk hogyan befolyásolta őket a várttól eltérő ejtésvariációk kezelésében.

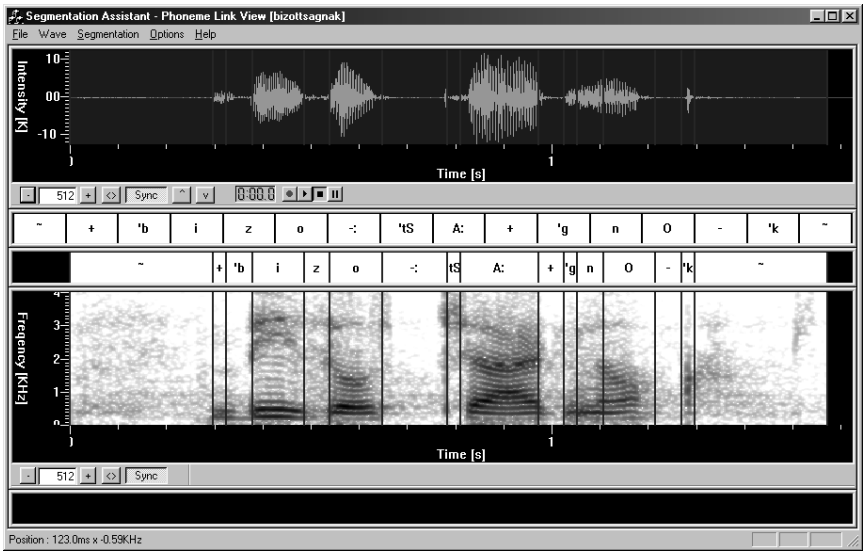
1. A szegmentálás

A statisztikai alapú beszédfeldolgozáshoz, különösen a gépi beszédfelismeréshez jól megszerkesztett, nagyméretű beszédadatbázisokra van szükség. A felismerők betanítása nem más, mint statisztikai alapú paraméterbecslés. A pontos becsléshez nagyszámú minta alapján megvalósuló betanítás szükséges. E minták gyűjteményei – a szükséges jegyzetekkel, címkézésekkel és átírásokkal – képezik az adatbázist. Az adatbázisoknak tartalmazniuk kell azokat a megfigyeléseket, amelyek a paraméterbecsléshez szükségesek; mindazokat a mintákat, amelyek egységesen lefedik a beszéd (és a környezeti zajok) változatosságát.

Az adatbázis legfontosabb része a szegmentált hanganyag. A szegmentálás a beszédfolyamat lineáris tagolása, vagyis a hangtest hangegységekre bontása. A hangtestben előforduló diszkrét metszetek kijelölésével jutunk el a beszédhanghoz, a hangtest legkisebb szegmentális szerkezeti egységéhez (Kiss, 2001). A beszéd időfüggvényében bejelöljük a fizikailag megfigyelhető beszédhangokat és azok határait. A szegmentálás célja, hogy az adatbázis gépi feldolgozásához megadjuk a beszédjel és a fonetikai átirat közti időbeli kapcsolatot, azt, hogy melyik szimbólum a beszédjel mely időintervallumának felel meg. A szegmentálás egységei a beszédhangok, ezekből absztraháljuk a fonémákat (Vicsi, *et al.*, 2002).

1.1 Fonetikai szintű szegmentálás és címkézés

A fonetikai szintű szegmentálás és címkézés célja kézileg, vagy esetleg egy automatikus szegmentálórutin támogatásával a beszéd időfüggvényében a fizikailag megfigyelhető beszédhangoknak és azok határainak a bejelölése. Ezt az ún. „audiovizuális fonetikai átírást” az 1997-ban elkészült BABEL nemzetközi project ajánlása alapján (Vicsi, Víg, 1998) érdemes végezni. Az átírásban a szöveg lehallgatása, továbbá az időfüggvény és/vagy a színkép elemzése nyújt segítséget. Ezen elemzések elvégzéséhez készítettünk egy speciális célprogramot, amelynek felületét az 1. ábra mutatja:



1. ábra: A szegmentálóprogram kezelői felülete

A program felső panelja a beszédjel hullámképét mutatja, az alsó panel pedig a színképelemzés révén előálló ún. spektrogramot. A lehallgatás mellett ez a két vizuális információ segíti a szegmentálást. Középen látható a fonetikus szimbólumok sorozata, amelyet hozzá kell rendelni az adott hangfelvételhez. Ez a fonetikus átirat természetesen javítható is, ha a beszélő esetleg mást mondott, mint amit a szoftver feltételezett. A fonetikus jelek és a hangjel egyes részleteinek összerendelése a határvonalak segítségével történik, amelyek elhelyezése után a jelek automatikusan a megfelelő hangrészlet fölé ugranak így segítve a tájékozódást.

2. Automatikus szegmentálás

2.1 Automatikus szegmentálás kényszerített illesztéssel

A beszédhang szintű szegmentálást és címkézést csakis nagy figyelmet igénylő és hosszadalmas kézi munkával lehet elvégezni. Megkönnyítheti és meggyorsíthatja viszont a munkát egy megfelelő, speciálisan erre a célra kialakított algoritmus, amely megkísérli automatikusan elhelyezni a fonetikai határokat. Bár ezt a feladatot jelenlegi tudásunk szerint tökéletesen nem tudjuk megoldani, olyan program azért készíthető, amely a határok jó részét elég pontosan helyezi el. A szegmentáló személy feladata ilyenkor csak az automata szegmentáló javaslatainak ellenőrzése és korrekciója, ami által a szegmentálás felgyorsítható.

Az automata szegmentálási algoritmusok közül a legjobbak azok, amelyek gépi tanuláson alapulnak. Speciálisan, a gépi beszédfelismerők is felhasználhatók a szegmentálási feladat elvégzésére. A beszédfelismerő algoritmusok ugyanis egy mondat felismerése közben keresést végeznek: a felismerendő hangjelre megpróbálják ráilleszteni az összes lehetséges fonetikai átiratot. Mivel a felismerés során a beszédhangok határai sem ismertek, ezért a felismerők végigpróbálják az összes lehetséges szegmentálást is. A felismerés eredményeképpen azt az átiratot, illetve szegmentumhatár-sorozatot adják vissza, amelyet az adott jel esetén a legvalószínűbbnek találtak. Egy beszédfelismerési alkalmazás esetén persze nincs szükségünk a szegmentumhatárokra, így az eredménynek ezt a részét figyelmen kívül hagyjuk. Viszont a felismerőnek ez az amúgy rejtve maradó „képessége” remekül kihasználható az automatikus szegmentálás céljaira. Ilyenkor ráadásul könnyebb a feladat, mint a felismerés esetén, ugyanis a fonetikai átirat adott, így azt nem is kell keresni; csupán az

adott átirat és a jel közötti optimális beszédhanghatár-összerendelést kell megtalálni. A beszédfelismerők ilyesfajta felhasználását „forced alignment”-nek nevezi a szakirodalom.

Eddigi munkáink alapján mi is készítettünk egy automata szegmentálórutint a fent leírt módon beszédfelismerő felhasználásával. A beszédfelismerő az MTA-SZTE Mesterséges Intelligencia Kutatócsoportnál fejlesztett „OASIS” rendszer volt, amely beszédhang alapú, és mesterséges neuronhálókat alkalmaz a beszédhang-felismerésre. A módszer technikai hátterét korábban már részletesen ismertettük (Zsigri, *et al.*, 2004), így attól itt most eltekintünk. A továbbiakban inkább arra térünk ki, hogy az automatikus szegmentáló használata során milyen nehézségek jelentkezhetnek, és ezeken hogyan lehet segíteni.

2.2 A kényszerített illesztés problémái

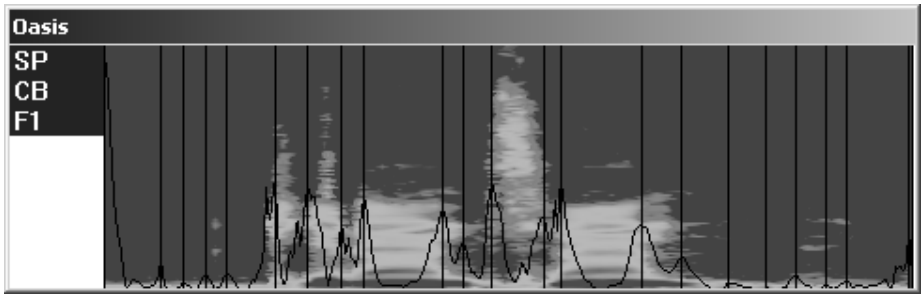
Habár jelenleg a beszédfelismerők „kényszerített illesztéses” felhasználása tűnik a legjobb automatikus szegmentálási módszernek, ez sem mentes a maga problémáitól. Ezek közül a legfontosabb, hogy az illesztés során használt fonetikus átirat esetenként lényegesen eltérhet a szegmentálandó beszédjel valódi tartalmától. A fonetikus átiratot ugyanis algoritmikusan állítjuk elő a beolvasott szöveg betű szerinti leírásából. Egy adott mondatnak pedig többféle eltérő kiejtése is lehet. Az alternatívák főleg a szóhatárok hasonulásainak végbemeneteléből vagy hiányából adódnak, de más tényezők is befolyásoló hatásúak, mint például a beszédsebesség vagy a beszélő artikulációs igényessége (vagy igénytelensége). Az ejtésvariációk kezelése nemcsak az automatikus szegmentálás esetén, hanem a beszédfelismerők szótárának automatikus kialakításánál is kulcsfontosságú, ezért vizsgálatuk máris megindult (Vicsi, *et al.*, 2003). A Műegyetemen fejlesztett fonetikus átírórendszer kiejtési opciókat is megenged a fonetikai átiratban, és az eredmények arra utalnak, hogy ez valóban valamivel jobb felismerési eredményekhez vezet (Mihajlik *et al.*, 2001). Azonban az MTBA adatbázis feldolgozása során nekünk nem állt rendelkezésünkre ilyen képességű fonetikai átíró, így az illesztések során csak egyfajta fonetikai átirat szolgált inputként. Ezzel gyakran előfordult, hogy bizonyos pontokon – például szóhatároknál – két-három egymást követő fonetikai szimbólumban sem felelt meg a beszédjelnek. Ilyenkor a beszédfelismerő rendszer kénytelen ráerőltetni a hibás szimbólumokat az adott jelszakaszra, illetve túl nagy eltérések esetén az algoritmus nem is hajlandó lefutni. Ha lefut is, ezeken a pontokon nyilván hibás eredménnyel. Ez pedig azt a veszélyt hordozza magában, hogy a későbbi manuális korrekció során a feldolgozó személy hajlamos átsiklani ezeken a részeken, és elfogadja a gép által kínált hibás megoldást.

2.3 Automatikus szegmentálás határtanulással

A beszédjelek automatikus úton történő beszédhangokra szegmentálása egyike a számítógépes beszédfeldolgozás klasszikus problémáinak. A tökéletes megoldásnak (amennyiben egyáltalán lehetséges) mindenképpen valamiféle tanulási módszeren kell alapulnia. Olyan technikák, amelyek pusztán jelfeldolgozáson alapulnak és az éppen adott szegmentálandó jelen kívül mást nem ismernek, nem tudnak teljes értékű megoldást szolgáltatni (ezt a megoldást ezért „félautomatának” neveztük, arra utalva, hogy a szegmentáló kimenetét utólag kézzel még korrigálni kell). Ezek az egyszerű módszerek a jel változását mérik, és a nagy változásoknál jeleznek feltételezett szegmentumhatárt. Korábbi munkánkban mi is kidolgoztunk egy olyan jelfeldolgozó függvényt, amely a szegmentumhatárokat hivatott jelezni (Zsigri, *et al.*, 2004). Ez a függvény a spektrum megfelelően kiválasztott sávjainak energiáit, pontosabban azok változását vizsgálja. Az egyes sávokra illesztett detektáló függvényeket súlyozott összegzéssel kombináltuk, ahol a megfelelő súlyértékeket tapasztalati úton, hosszas kísérletezéssel lőttük be. Ez természetesen nem volt optimális, ráadásul az egyik adatbázison jónak talált értékek nem feltétlenül

működtek egy másikon. Ezen javítandó megkíséreltük a paraméterek optimális értékét gépi tanulással megtalálni. Ez a következőképpen történt:

Tanító adatbázisnak az adatbázis kézileg már felszegmentált részét használtuk. Elvileg a határként bejelölt időpontok beszédkereteit kellett volna pozitív tanulópéldaként használnunk, a határok közé eső kereteket pedig negatív példaként. Ezzel az egyik probléma az, hogy ily módon jóval több lett volna a negatív példa, mint a pozitív, korábbi tapasztalataink szerint pedig ez gondot okozhat a tanulásnál. A másik nehézség pedig, hogy maguk a kézileg behúzott határok is lehetnek némileg pontatlanok. Ezért azt a megoldást választottuk, hogy a határok közé egy x^6 jellegű görbét illesztettünk oly módon, hogy az a két határnál vegye fel az 1-es értéket, a szegmentum középpontjánál pedig a 0-át. Ily módon a „határság" valószínűségét kissé „szétkentük", a kézileg behúzott határoktól való távolság függvényében. Az így előállt célfüggvény tanulására (regressziójára) egy kétrétegű előrecsatolt neuronhálót alkalmaztunk 20 rejtett neuronnal. Input jellemzőkészletként a kézi szegmentáló kifejlesztésekor kikísérletezett, sávenergiákból származtatott görbék szolgáltak. A tanulás után a neuronháló által szolgáltatott görbét szemlélteti a 2. ábra.



2. ábra: A *'kilencven'* szó spektrogramja, a neuronháló által szolgáltatott görbe, és az ez alapján algoritmikusan kiválasztott szegmenshatárok

A neuronháló kimenetéből legegyszerűbben úgy kaphatunk szegmentumhatárokat, ha megkeressük a lokális maximumhelyeit. Továbbá a görbe adott pontban felvett értéke valószínűségi értékként értelmezhető, így egy egyszerű küszöböléssel szabályozhatjuk a behúzott határok számát, ekképp egyensúlyozva a törlési és beszúrási hibák között. A 2. ábrán bejelölt határokat is a lokális maximumhelyekből kiindulva állítottuk elő. Tapasztalataink szerint a neuronhálós megoldás sokkal stabilabban működik, mint a korábbi, empirikus úton belőtt félautomata szegmentálónk, és bizonyos fajta hanghatárokat nagyon jó pontossággal talál el. Hibákat főleg olyan esetekben követ el, amely esetek a kézi szegmentálás során is a legnehezebbeknek bizonyulnak (pl. magánhangzó-félmagánhangzó átmenet). Ha a pontos fonetikus átirat adott, akkor természetesen ez a fajta szegmentálás rosszabb eredményt ad, mint a kényszerített illesztés, hiszen a fonetikus átirat ismerete nélkül kell dolgoznia. De ha a fonetikus átirat komoly hibákat tartalmaz, akár jobbnak is bizonyulhat. Továbbá a küszöbérték beállításával elérhető, hogy az algoritmus csak a biztos helyekre húzzon, a több határt hagyja ki. Ilyenkor a kézi ellenőrzést és korrekciót végző személy kénytelen elhelyezni a további hiányzó határokat, míg a kényszerített illesztéssel dolgozó algoritmus mindig a szükséges számú határt húzza be, amelynek hibáin a figyelmetlen szemlélő esetleg átsiklik.

2.4 Határtanulás a beszédfelismerésben

Az előző fejezetben ismertetett határtanulási módszer a beszédfelismerésben is hasznos lehet. Mint korábban említettük, futásuk során a felismerők gyakorlatilag végigvizsgálják az összes lehetséges fonetikai szegmentálást. Ez egy óriási hipotézistér bejárását jelenti, nyilván komoly időigénnyel. Amennyiben bizonyos szegmentálási

lehetőségeket egy gyors algoritmussal ki tudunk zárni, akkor ezeket az eseteket nem kell a felismerőnek kiértékelnie, így gyorsabb működést érhetünk el. Természetesen mindig fennáll a kockázata annak, hogy a fürdővízzel a gyereket is kiöntjük, azaz véletlenül egy valódi határt is kidobunk. Ezt a felismerő többnyire már nem képes korrigálni, ezért inkább túl sok, mint túl kevés határt kell behúznunk. Az előző fejezetben ismertetett algoritmus használata esetén egyszerűen a küszöbérték módosításával szabályozhatjuk, hogy hány határ maradjon meg.

Ha számszerűsíteni akarjuk egy automatikus szegmentáló hatásfokát, akkor legjobb, ha összevetjük a kézi szegmentálással. Erre a következő, "edit distance" jellegű algoritmust alkalmaztuk: a gépi és a kézi határsorozat határait összepárosítjuk úgy, hogy a párok távolságainak összértéke minimális legyen (ez dinamikus programozással megoldható). Ha a két határsorozat nem ugyanannyi határból áll, akkor nyilván lesznek kimaradó határok. Nem fogadjuk el tovább azokat a párokat, amelyeknél a távolság egy ezredmásodpercben adott küszöbnél nagyobb. Végeredményként a kézi szegmentálás pár nélkül maradó határainak száma adja az algoritmus törlési hibáinak számát, a fölöslegesen behúzott (azaz pár nélkül maradt) gépi határok száma pedig a beszúrási hibák számát.

Az algoritmus tesztelését az OASIS-Numbers adatbázis (http://www.inf.u-szeged.hu/oasis) egy részén végeztük el, amely jó minőségben rögzített számbemondásokat tartalmaz. Azt találtuk, hogy 30 ezredmásodperces hibaküszöb mellett a 10488 valódi szegmenshatárra 28262 beszúrási és 88 törlési hiba jutott. Ez azt jelenti, hogy az algoritmus kb. négyszer annyi határt húz be, mint kellene, viszont a kihagyások száma egy százalék alatt van! A négyszeres „határtúladagolás” soknak tűnhet, de valójában ez csak töredéke annak, ahány lehetőséget a felismerő végigvizsgál, így a szegmentáló használata látványos gyorsulást eredményezett a felismerő futásában. Meg kell jegyeznünk azonban, hogy az OASIS-Numbers adatbázis csak számbemondásokat tartalmaz, amelyeket egyrészt elég gondos artikuláció jellemez, másrészt rengeteg hangkapcsolat egyáltalán nem fordul elő bennük. Ezért az algoritmus további tesztelését tervezzük az MTBA adatbázison.

3. A kézi szegmentálás elvei és tapasztalatai

Az automata és a félautomata szegmentálás ígéretes eredményei ellenére kézi munkával érhetünk el igazán pontos fonéma szintű szegmentálást, annak ellenére, hogy ez igen fáradságos és időigényes munka.

3.1 A szegmentálás szempontjai

A beszédhangok a lineáris hangszerkezetben hangkapcsolatokat képeznek. A hangok gyakran nem különíthetők el élesen egymástól, rövidebbek-hosszabbak köztük az átmenetek, ezért a munka során azt tapasztalhatjuk, hogy a beszédhanghatárt nem lehet mindig egyértelműen bejelölni. Gyakran a spektrogram és a hullámforma összehasonlítása, de még a meghallgatás sem ad objektív fogódzót a határ pontos bejelölésére. Ezekre az esetekre az MTBA adatbázis feldolgozása elkezdése előtt az alábbi elveket dolgoztuk ki (Vicsi, *et al., 2002*):

- A szegmentumhatárokat érdemes a nullátmenetekhez igazítani.
- Zöngés hangok esetében ez a pozitív null átmenetet jelenti. Nagyon precízen kell jelölni a határt. Zöngétlen hangoknál 1 ms pontossággal lehet jelölni a hang kezdetét.
- A magánhangzó kezdetét a zönge indulásánál kell jelölni (zöngétlen hang után).
- A zárhangok, affrikáták kezdetét a megelőző hang utolsó periódusa előtt jelöljük.
- A magánhangzó-magánhangzó vagy magánhangzó-rezonáns mássalhangzó kapcsolatokban a határt az átmeneti rész 50%-ánál jelöljük be. Ilyenkor némileg pontatlan a bejelölés, mert a hangok kettéválasztása bizonytalan (Vicsi, Víg, 1998).

- Előfordulhat, hogy egy-egy hangot többszöri visszahallgatás után sem lehetett azonosítani. Ennek jelzésére a [cut] kódot lehet használni.

3.2. Szegmentálási tapasztalatok

Az MTBA adatbázis kézi szegmentálásának elkészülte után igyekeztünk összegyűjteni a tapasztalatainkat. Egyik szempontunk az volt, hogy mennyiben sikerült betartani az előzetesen rögzített szegmentálási alapelveket. A másik nézőpontunk az volt, hogy a szegmentálók hogyan kezelték az ejtésvariációkat, azaz mihez kezdtek olyankor, ha a gép által javasolt fonetikai átirat eltért az általuk érzékelttől. Ezekből fontos tanulságok vonhatók le az automatikus átirat, illetve a felismerők szótárának összeállítására nézve, így főleg az ilyen jellegű tapasztalatainkat összegezzük az alábbiakban.

A kézi szegmentálást végzők az előző pontban felsorolt elvek közül az egyiktől rendszeresen eltértek, anélkül, hogy ezt észrevették volna. Ez az elv a következő volt:

- A magánhangzó kezdetét a zönge indulásánál kell jelölni (zöngétlen hang után).

Ez az angolban vagy németben jól alkalmazható elv a magyarban csak a réshangok utáni magánhangzók kezdetének a kijelölésére használható. Az angolban vagy a németben zárhangok esetén is jól működik, mert ezekben a nyelvekben a zönge jóval a zár felpattanása után kezdődik.[1] A magyarban viszont a felpattanáskor szinte azonnal rezegni kezdenek a hangszalagok. Ha a szegmentálók valóban a zönge indulásánál húzták volna be a vonalat, akkor a zárhangból nem sok maradt volna. Szerencsére nem ezt tették, hanem a fülükre hallgatva a felpattanó zörejt meghagyták a zárhangnak.

A szegmentálási szempontok közül az, hogy a zárhangok, affrikáták kezdetét a megelőző hang utolsó periódusa előtt jelöljük, újdonság a korábbi gyakorlathoz képest. Korábban ezt a szakaszt az egyszerűség kedvéért a megelőző hanghoz soroltuk. A határvonal balra tolását az indokolja, hogy így a fel nem pattanó zárhangok felismerése is lehetővé válik (Sejtes – Zsigri, 2003). Ezeknek az aránya ugyan nem jelentős a magyarban, de annyi azért van belőlük, hogy ne mondjunk le róluk. Leggyakrabban azonos képzési helyű orrhangok előtt maradhat el a zár felpattanása, pl. *népmese, kötni,* vagy hasonulásos alakokban: *pillanatnyi, vadnyúl,* de néha megnyilatkozás végén is megfigyelhető.

A beszédfelismerési folyamatnak abban a szakaszában, amelynek a végén eljutunk a beszédhangokhoz, általában még nincs szükség szótárra. Tisztán statisztikai alapon is hozzá lehet rendelni hangszakasztípusokhoz beszédhang-szimbólumokat. De már itt is vannak nem egyértelmű esetek.

A szótagvégi *l* gyakran eltűnik, és megnyújtja az előtte levő magánhangzót. A megnyúlt magánhangzó nem azonos a rövid magánhangzó hosszú párjával: minőségében ugyanolyan marad, mint a rövid magánhangzó, csak az időtartama lesz hosszabb, pl. *elment* > *ēment* [E:mEnt], és nem *ément* [e:mEnt] (Nádasdy – Siptár, 1994). Ilyen esetekben a szegmentáló hiába keresi az *l*-et, mert az nincs ott. Jó beszédfelismerést valószínűleg az eredményezhet, ha a szegmentáló nem próbálja önkényesen rövid [E]-re és [l]-re bontani a hosszú [E:]-t, hanem a valósághoz hűen csak egy szegmentumot vesz fel, és ezt [E:]-vel jelöli. Ezt a beszédfelismerő program olyan bemeneti adatként kezeli, amelynek az egyik lehetséges kimenete az /E/+/l/ fonémakapcsolat (illetve <el> betűkapcsolat). Azért csak az egyik lehetséges kimenete, mert fizikailag ez az [E:] nem különíthető el az egymást szünet nélkül követő két [E]-től, pl. *leesett*.[2] A szegmentálók ugyan mindig két szegmentumként elemzik a *leesett*-nek az *l* és *s* közötti szakaszát, de a beszédfelismerő program nem mindig kap kézzel szegmentált anyagot. Kézi szegmentálás híján viszont pusztán a hullámformából lehetetlen

[1] A felpattanás és a zöngeindulás közötti zöngétlen szakaszt nevezik hehezetnek.

[2] Budapesti köznyelvi adat. A kétféle *e*-t ismerő nyelvváltozatokban a *leesett* két *e*-je nem egyforma

eldöntenie, hogy az elemzendő hangszakasz az /El/ vagy az /EE/ fonémakapcsolat megvalósulása-e. Ez csak szótár segítségével lehetséges. Ha a szótárban benne van, vagy a szótár elemeiből kialakítható az, hogy *leesett*, de az nem, hogy *lelsett*, akkor az [E:] ~ [EE] bemeneti adat /E]/+/E/ kapcsolatként interpretálandó.

A szegmentáló embernek megvan az az előnye a géppel szemben, hogy ismeri a nyelvet, és előre el tudja dönteni a többféleképp is interpretálható hangszakaszokról, hogy melyik a helyes interpretáció. Valójában nem is szoktunk tudatában lenni annak, hogy amit elemzünk, azt másképp is lehetne elemezni. A *kiutazik* [k]-ja és [t]-je közötti szakaszt mindenki két szegmentumra bontja ([i]-re és [u]-ra), a *kijut* [k]-ja és [t]-je közötti szakaszt pedig háromra ([i]-re, [j]-re és [u]-ra). De csak akkor, ha az egész szót hallják. Ha kivágjuk a *kiutazik* „iu"-ként értelmezett szakaszát és a *kijut* „iju"-ját, és a szegmentálóknak csak a kivágott szakaszt adjuk oda, rögtön nem lesz egyértelmű, hogy melyik szakasz hány szegmentumból áll. Ennek az az oka, hogy az [i]-t szünet nélkül követő magánhangzók olyankor is [j]-vel kapcsolódnak az [i]-hez, ha azt a helyesírás nem jelöli, pl. *fi*[j]*am*, *Pistá*[j]*ék*. A jó helyesíró szegmentálók, ahol nem írnak <j>-t, ott nem is keresnek [j]-t. Az automata szegmentáló viszont csak úgy mehet biztosra, ha a kétféleképp is értelmezhető hangszakaszokról szótár segítségével dönti el, hogy melyik értelmezést fogadja el.

A hasonult és összeolvadt alakok visszaalakítása már a beszédhangok megállapításával záruló szakasz után következik, amelyben a szótár még fontosabb szerepet kap. Azt, hogy a *pénzt* szóban a [t] előtti zöngétlen beszédhang egy zöngés fonémának a megvalósulása, és emiatt nem <sz>-szel írandó, hanem <z>-vel, vagy hogy az *aludj*, *hagyj* és *higgy* hosszú [d':]-je csak kiejtésben azonos, leírva mind a három más, csak szótár segítségével állapítható meg. Ez a szótár természetesen többféleképp is létrehozható: begépeltethető emberekkel is, de a gép is kialakíthat magának egy olyan, az emberi szótáraktól esetleg jelentősen eltérő szótárszerű képződményt, amelynek alapján megtippelheti, hogy egy beszédhanghoz mikor milyen betű vagy betűkapcsolat rendelhető hozzá a legvalószínűbben. A szótárban való keresés hatékonyságát növelheti, ha beszédfelismerő program kellő „tapasztalatokkal" rendelkezik arról, hogy melyik beszédhangot milyen betű jelölheti. Ezek a tapasztalatok hosszadalmas betanítással nyelvész közreműködése nélkül is megszerezhetők, de egy jól algoritmizálható hangtani leírás valószínűleg jelentősen csökkentheti a betanítási időt.

A programtervező matematikusok és nyelvészek együttműködése nem mindig egyirányú. Bár az esetek többségében a programtervezők építik be saját rendszereikbe a nyelvészet által felhalmozott ismereteket, időnként vissza is fordul ez a folyamat. Az /l/ fonémának egy olyan allofónja, amelyről a magyar nyelvű szakirodalomban nem írnak, egy beszédfelismerő program betanítása során jelentkezett olyan gyakorisággal, hogy feltűnt a szegmentálóknak. Arról, hogy a *-tlan/-tlen* képzőben nagyon gyakran zöngétlen az /l/, (Tóth, *et al.*, 2003)-ban olvashatunk először.

Irodalom

Barry, W.J. and A.J. Fourcin „Levels of labelling", Computer Speech and Language, Vol. 6 1992, ,pp. 1-14.

Duda, R. O., Hart, P. E., Stork, D. G.: *Pattern Classification*, Wiley and Sons, 2001

Huang, X., Acero, A., Hon, H.-W.: *Spoken Language Processing*, Prentice Hall, 2001

Kiss J.: *Magyar dialektológia*. Budapest, Osiris Kiadó, 2001

Mihajlik, P. és Tatai, P.: Automatikus fonetikus átírás magyar nyelvű beszédfelismeréshez, Beszédkutatás 2001.

Nádasdy Á., Siptár P.: A magánhangzók, in. Kiefer F. (szerk.): *Strukturális magyar nyelvtan* 2: *Fonológia*. Budapest: Akadémiai Kiadó, 42–182.

Pollak, P., Cernocky, J., Boudy, J., Choukri, K., Heuvel, H., Vicsi, K., Virag, A., Siemund, R., Majewski, W., Sadowski, J., Staroniewicz, P., Tropf, H., Kochanina, J.,

Ostrouchov, A., Rusko, M., Trnka, M.(29 May 2000): SpeechDat(E) –Eastern European Telephone Speech Databases. Proceeding LREC' Satelite workshop XLDB – Very large Telephone Speech Databases , Athens, 2000

Sejtes Gy., Zsigri Gy.: Hangátmenetek a beszédfelismerésben, in: Alexin Z., Csendes D. (eds.): Magyar Számítógépes nyelvészeti Konferencia 2003, Szeged, pp. 176–181.

Tóth, L., Kocsor, A.: Az MTBA magyar telefonbeszéd-adatbázis kézi feldolgozásának tapasztalatai, Beszédkutatás 2003, pp. 134-146.

Vértes, O. A.: *Bevezetés a fonetikába*. Második, bővített kiadás. Budapest: Gyógypedagógiai Tanárképző Főiskola, 1952

Vértes, O. A.: Az artikuláció akusztikus vetülete, in Bolla Kálmán szerk. *Fejezetek a magyar leíró hangtanból*, 155–164. Budapest: Akadémiai Kiadó, 1982

Vicsi, K., Tóth, L., Kocsor, A., Gordos, G., Csirik, J. MTBA-Magyar nyelvű telefonbeszéd-adatbázis, Híradástechnika, LVII. 2002/8, Budapest, pp. 35-43.

Vicsi K., Vig A.: Az első magyarnyelvű beszédadatbázis, Beszédkutatás '98, MTA Nyelvtudományi Intézete, Budapest 1998, pp. 163-177

Vicsi, K., Szaszák, Gy.: A magyar nyelv kiejtésvariációi és felhasználásuk a beszédfelismerésben II., Beszédkutatás 2003, pp. 163-176

Wells, J. at all.: Standard Computer-Compatible Transcription. Esprit Project 2589 (SAM), Doc. no. SAM-UCL-037. London: Phonetics and Linguistics Dept., UCL, 1992

Zsigri, Gy., Kocsor, A., Tóth, L. és Sejtes, Gy.: Phonetic Level Annotation and Segmentation of Hungarian Speech Databases, accepted for Acta Cybernetica