Beszédadatbázisok előkészítése
kutatási és fejlesztési célok hatékonyabb támogatására

NÉMETH GÉZA, OLASZY GÁBOR,
BARTALIS MÁTYÁS, ZAINKÓ CSABA, FÉK MÁRK, MIHAJLIK PÉTER

BME Távközlési és MédiainformatikaiTanszék
{nemeth, olaszy, bartalis, zainko, fek, mihajlik}@tmit.bme.hu

Lektorált

Kulcsszavak: beszédadatbázisok, címkézés, korpusz alapú beszédszintézis, hanghatárkorrekció

A nagyméretű beszédadatbázisok készítése az utóbbi évtizedekben vált szükségessé, hogy támogassák egyrészről a beszédkutatást, másrészről a működő beszédinformációs rendszerek fejlesztését. Az ilyen adatbázisok akkor szolgálhatják jól a tudományt, ha részletes belső címkézéssel is rendelkeznek. Jelen cikkben olyan adatbázisokkal foglalkozunk, amelyek egyetlen bemondótól felvett, sok mondatból álló, több órányi anyagot tartalmaznak. Az ilyen beszédadatbázisok címkézésénél alapvető gond, hogy a címkézést teljes mértékben emberi erővel nem lehet elvégezni a munka nagysága miatt. A cél viszont az, hogy a címkék a lehető legpontosabban legyenek bejelölve a hullámformában. A cikkben ismertetett új hibrid eljárást eredményesen lehet alkalmazni az ilyen munkákhoz, szinte hibamentes címkézés érhető el, időigénye is elviselhető (2-3 nap egy több órás adatbázisra). Az így készített beszédadatbázisokban a keresés megbízható eredményeket ad, melyet fel lehet használni a beszédkutatásban, a beszédszintézisben és a beszédfelismerésben is.

1. Bevezetés

A részletes címkézés érintheti a szegmentális szerkezetet (hanghatárok, szavak határai), valamint a szupraszegmentális szintet (hangsúlyok, dallammenetek, szünetek, prozódiai egységek). Az adatbázisok címkézési munkáit nagy méretük miatt csak jelentős szoftver-támogatással lehet költséghatékonyan elvégezni. Léteznek már évek óta magyar beszédadatbázisok, amelyeket főleg beszédfelismerő algoritmusok tanítására fejlesztettek [7,8]. Ezekben általában sok beszélőtől vettek beszédmintát és a címkézési munkákat még többnyire jelentős mértékben kézi erővel végezték.

Jelen cikkben olyan adatbázisokkal foglalkozunk, amelyek egyetlen bemondótól felvett, sok mondatból álló anyagot tartalmaznak. Ezek címkézéséről van szó. Egyelőre a hanghatárok bejelölésével kapcsolatos szoftverrendszer fejlesztéséről és annak működési tapasztalatairól számolunk be. A rendszert a BME Távközlési és Médiainformatikai tanszékén fejlesztették és az ottani beszédadatbázisokhoz használják. Az eljárás fontos tulajdonsága, hogy szoftveres elemek és emberi erőforrás váltogatják egymást a feldolgozás során. A beszédfeldolgozás egyes pontjain még ma sem lehet kihagyni az emberi döntéshozatali tényezőt.

Az eredmények igazolják, hogy ilyen hibrid eljárással elérhető a szinte hibamentes címkézés, ennek ára viszont a bonyolult, kissé időigényesebb feldolgozás. Az ilyen adatbázisokból pontos és megbízható adatok nyerhetők. A vizsgált adatbázisokról kapott információk azt is megmutatják, hogy az egyes beszélők közötti hangszintű beszédképzési eltérések számszerű adatokkal is jellemezhetők, ami a személyre szabott szoftveres beszédjellemzés egyik kísérleti megvalósításának is tekinthető.

2. A munka célja, módszere és a feldolgozott anyag

A nagyméretű beszédadatbázisok hullámformáját ma már el lehet látni hanghatár-címkékkel szoftver segítségével, azonban az eredmény sohasem teljesen pontos. Ez annak a következménye, hogy a beszédjel biológiai mechanizmus terméke. A beszéd előállítása során a pillanatnyi motoros és artikulációs történések határozzák meg a kisugárzott hullámformát (ugyanazon mondat többszöri kiejtése során minden esetben más akusztikai eredményt kapunk, a hangzás csak globálisan, nyelvi szempontból lesz ugyanaz). A gépi címkézés pontossága sok tényezőtől függ. A jelen kísérlet célkitűzése az, hogy tegyük teljessé a címkézést, a szoftveres alapcímkézés eredményét javítsuk tovább célzott szubrutinokkal, a géppel nem javítható hibákat pedig emberi erővel javítsuk.

Eredményként olyan beszédadatbázist kapunk, amelyikben egyrészről úgymond minden hanghatár helyesen van bejelölve, másrészről a beszédhangok minőségi osztályozására is vannak jelzések. Ez utóbbi megjegyzésen a következőket értjük. Az emberi beszédben a hangkapcsolódások artikulációjából adódóan előfordulhatnak olyan hangok, amelyek belső szerkezetüket tekintve torzultak és nem felelnek meg a fonetikai hangleírásoknak [4]. Ezek a torzult hangok ugyan szerves részét képezik a hangsornak, de csak a saját szélesebb hangkörnyezetüket tekintve (a szó, amiben szerepelnek) adják meg az emberi percepciós rendszernek a megértéshez szükséges akusztikai információkat. Az ilyen hangok megjelölése azért fontos, mert az adatbázis felhasználása során ezek a hangok szétválaszthatók a többitől (például hangzásvariációk keresésekor, vagy beszédszintézisnél, amikor el kell dönteni, hogy mikor melyik hangot használjuk

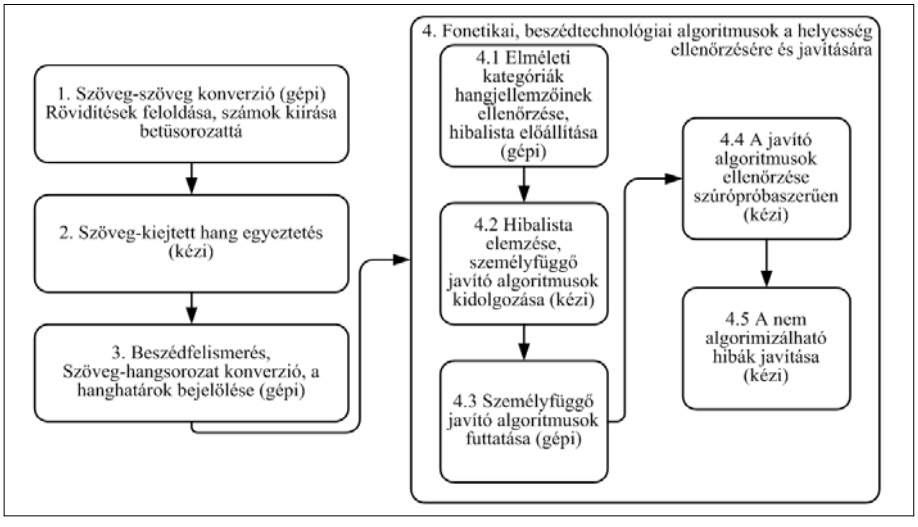
stb.). Az ilyen, jól címkézett beszédadatbázisok a későbbiekben sokféle célra felhasználhatók az oktatásban, a kutatásban és az alkalmazás-fejlesztésekben is.

A cél megvalósítására a hagyományos egylépcsős gépi megoldással szemben fokozatos kialakítású, többlépcsős javító rendszert fejlesztettünk ki. A kísérletek azt mutatták, hogy a nagy pontosságú címkézéshez olyan hibrid megoldást kell keresni, amelyben az egyes lépcsők futtatása között emberi döntéseket is kell hozni, hangolni kell bizonyos szoftverelemeket (nem teljesen automatikus címkézésről van szó). A célkitűzéshez tartozott az is, hogy minimumra csökkentsük a manuálisan javítandó hibák számát, vagyis ésszerűen kézben tartható hibamennyiség maradjon a szoftveres feldolgozás után (max. 500 hiba/adatbázis, amely manuálisan 2-3 óra munkával javítható). A feldolgozáshoz és az eredmények megjelenítéséhez a Praat 4.0 szoftvert is használtuk [1].

A gyakorlati tapasztalatok azt is megmutatták, hogy a címkézési folyamat beszélőfüggő megoldást kíván egészen addig, amíg elegendő mennyiségű, különböző beszélőtől felvett adatbázis feldolgozása meg nem történik (az adott nyelvre). Az ilyen adatbázisok készítésénél maga a hangfelvétel létrehozása is komoly munkát igényel, ezért jelenleg kevés beszélőtől áll rendelkezésre nagyméretű beszédadatbázis. A jelen munkában csak a kezdetekről tudunk beszámolni, mindössze négy beszédadatbázis hangfelvételét készítettük el és dolgoztuk fel (három képzett női beszélő: N1, N2, N3 és egy férfihang: F1). Megvizsgáltuk a beszélők közötti kiejtési (hangképzési) jellegzetességeket is, amelyek befolyásolják a helyes szoftveres címkézést. Mindegyik adatbázis a BME Távközlési és Médiainformatikai Tanszék beszédkutató laboratóriumában található. Az adatbázisok néhány jellemző adatát az *1. táblázatban* láthatjuk.

3. A címkézésre kidolgozott módszer

A új, hibrid gépi címkézési eljárás moduljai az *1. ábrán* láthatók.



1. ábra
A beszédadatbázis hanghatárainak címkézési folyamata

4. Az alapcímkézés

Az alapcímkézést géppel végezzük (4.3. pont), a hibajavításokat gépi és emberi feldolgozással. A hangfelvétel elvégzése után két előkészítő lépést kell elvégezni, hogy a gépi alap címkézést elkezdhessük.

4.1. Első lépés

A felolvasáshoz alkalmazott karaktersorozatot betűsorozattá alakítjuk, ahol a szövegben rövidítés vagy szám van, azt feloldjuk és betűkkel kiírjuk. Az így átalakított eredeti szöveg csak betűkaraktereket tartalmaz majd. A konvertálást célprogram végzi, amely a Profivox beszédszintetizátor szöveg-szöveg átalakító moduljának felhasználásával készült [6].

Példa:
Eredeti szöveg: A hőmérséklet június 3-án Bp-en -3 C° körül várható.
Átirat: A hőmérséklet június harmadikán Budapesten mínusz három celziusz fok körül várható.

4.2. Második lépés

Ebben a lépésben végezzük el a szöveg-beszédhullám szinkron ellenőrzését. Egy szakértő meghallgatja a mondatokat, közben vizuálisan ellenőrzi a hozzájuk tartozó szöveget (a 4.1. szerint). E vizsgálattól függ a további munka sikeressége. Ezt a folyamatot csak ember tudja elvégezni. Ez a munkafázis meglehetősen sok időt vesz igénybe és komoly koncentrációt kíván. A talált hibák kijavítása után elméletileg a két médium szinkronban van egymással (a gyakorlat azt mutatja, hogy néhány hiba azért benne marad valamelyikben, ez a későbbi szoftverellenőrzéskor kiderül és akkor javítjuk).

1. táblázat
A feldolgozott beszédadatbázisok adatai

Adatbázis	Mondat-szám	Szavak száma	Hangok száma	Hangzási idő (perc)	Beszédseb. hang/s	Artik. seb. hang/s	Megjegyzés
N1	5821	102940	488093	641,7	12,89	13,28	időjárásjelentés
N2	3643	43345	259353	331,6	13,30	13,44	telefon árlista
N3	792	9086	39353	50,3	12,68	12,71	prompt
F1	430	3071	11822	15,3	12,66	12,66	általános szöveg

4.3. Harmadik lépés

A bevezető két lépés után következik az alapvető gépi hangfelismerés és címkézés (a beszédhangoknak és határaiknak a meghatározása). Szükség van az adott hangfelvételre, annak ortografikus szöveges átiratára és az alkalmazandó (opcionálisan vagy kötelezően) végbemenő hasonulási, összeolvadási stb. jelenségek megadására. Az úgynevezett kényszerített illesztési (forced alignment) üzemmódú beszédfelismerő első lépésként a csak betűkből álló szöveget (a 4.2.-ből) hangszimbólumok sorozatává alakítja. Ez a szimbólumsorozat fogja segíteni a hanghatárok felismerését (a gép tudja, hogy milyen hang lehet az adott ponton). A hangszimbólumokat itt most két ferde zárójel közé írva adjuk meg.

Példa:
A hőmérséklet június harmadikán Budapesten mínusz három celziusz fok körül várható.
/a/ /h/ /o3/ /m/ /e1/ /r/ /s/ /e1/ /k/ /l/ /e/ /t/ /j/ /u1/ /n/ /i/ /u/ /s/ /h/ /a/ /r/ /m/ /a/ /d/ /i/ /k/ /a1/ /m/ /b/ /u/ /d/ /a/ /p/ /e/ /s/ /t/ /e/ /m:/ /i1/ /n/ /u/ /sz/ /h/ /a1/ /r/ /o/ /m/ /c/ /e/ /l/ /z/ /i/ /u/ /sz/ /f/ /o/ /k:/ /o2/ /r/ /u2/ /l/ /v/ /a1/ /r/ /h/ /a/ /t/ /o1/.

A hangszimbólumok a hozzájuk tartozó betűkkel megegyezők, kivéve az ékezetes betűket, amelyek alapját a főkarakter és az ékezetnek megfelelő szám kombinációja adja. Például: o=o, o1=ó, o2=ö, o3=ő. A hosszú hang jelölésére a kettőspontot használjuk.

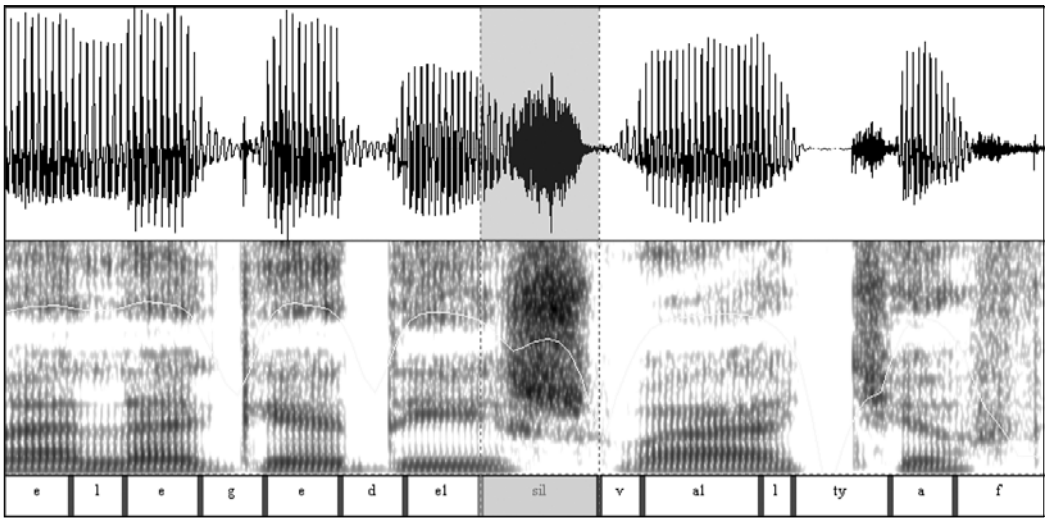
A hangszimbólum sorozat alapján a kényszerített illesztést alkalmazva – amikor is a felismerési hálózat a szavak lineáris szekvenciájából adódik – a felismerő egyszeri futtatásával előállnak mind a hanghatárok, mind a hang-identitások. Az eljárás kezeli a szóhatárokon átívelő hasonulási jelenségeket és az ezekből adódó hangkieséseket is. További részletek a gépi címkézési algoritmusról a [2,3]-ban találhatók.

A megközelítés feltételezi, hogy a felismerés során használt, az akusztikus modellek betanításához használt sokbeszélős adatbázis címkézése pontos volt. Amennyiben ez nem áll fenn, a szisztematikus címkézési hibák továbbadódnak. Ezek kiküszöbölése speciális célú algoritmusokkal lehetséges. A statisztikai elvből is következik, hogy néhány hanghatár pontatlan lehet. Végül meg kell említeni, hogy mivel a beszédfelismerő telefonsávi beszéddel lett betanítva, a magas frekvenciakomponensekben gazdag hangok, mint a „c”, „sz” stb. határainak pontos felismerése elvileg is lehetetlen. Robusztussága miatt viszont bevált ez a megközelítés.

Fontos továbbá azt is megjegyezni, hogy ennek a hangfelismerési eljárásnak az a tulajdonsága, hogy nem vizsgálja sem a zöngés/zöngétlen állapotot, sem a hangidőtartamot, csak spektrális jellemzők alapján dönt. A program felismerési pontossága a fent felsorolt problémák ellenére nagyon jó, 95-96%-os. A hangfelismerés eredményeként megkapjuk a hanghatárok (címkék) sorozatát, amit a hullámformával párhuzamosan megjeleníthetünk *(2. ábra)*.

A megjelenítés alapján vizuálisan is ellenőrizhetjük a címkézés pontosságát (a 4.3. munkafázist többször is el kell végezni, ha a 3.1.2. pontban hibákat vét az ellenőrző személy). A vizuális ellenőrzést a későbbi hibadetektálásoknál és javításoknál is használjuk. Például a 2. ábrán a szürke sáv hibát jelez. A melegedés váltja fel hangsorban az s hang helyén szünet /sil/ jel szerepel a hangszimbólum reprezentációban. Ez egy szinkronizálási hibából fakad. Az eredeti szövegből kimaradt az s betű, így a felismerőnek eggyel kevesebb hangot kellett a hullámformába beilleszteni, ezért ide szünetet jelölt. További hiba, hogy az /e1/ magánhangzó végéből 4 periódus a zöngétlen /s/ hanghoz van jelölve.

A gépi felismerő tévesztései nagyban függnek attól is, hogy a beszélő személy beszédképzési mozzanatai mennyiben felelnek meg az elméleti fonetikai modell paramétereinek (például zöngésség/zöngétlenség, zárképzés, rövid-hosszú oppozíció stb.). Ha a beszélő gyorsan és laza artikulációval beszél, akkor több címkézési hiba lesz, ellenkező esetben kevesebb. Hibaforrás lehet továbbá a koartikulációból adódó olyan hangelem, amelyik fonemikus szinten nem köthető beszédhanghoz [5], illetve annak részéhez (svá, koartikulációs néma fázis, stb.) Ilyenkor a felismerő hibásan állapíthatja meg a hanghatárt. A harmadik hibacsoport, amikor a beszélőre jellemző hangképzési állapotok eltérnek az elméleti osztályozásoktól (például zöngétlen laterális mássalhang-



2. ábra
A hanghatárok és a hangszimbólumok (lent), valamint a hullámforma (fent) együttes megjelenítése a spektrális tartalommal (középen) együtt.

zó is kialakul, ilyen hangot nem definiál a magyar hangrendszer). Végül címkézési hibát okozhat a szinkrontól való eltérés is. Ilyenkor durva elcsúszások lehetnek (lásd a 2. ábrán). Szinkron hiba esetén a 4.2. pontra kell visszamenni a feldolgozásban, a szinkront helyre kell állítani és a felismertetést újból el kell végezni.

A további feldolgozás során tehát ezt a címkézett adatbázist (a hibáival együtt) tartjuk a következő, javító munkafázisok bemenetének.

5. Fonetikai alapú algoritmusok az ellenőrzésre és javításra

A célkitűzés az, hogy a gépileg helyenként hibásan bejelölt hanghatárokat (a teljes állomány 4-5%-a) a beszédadatbázisban feltárjuk és kijavítsuk. Javítási követelmény, hogy a hibák többségét algoritmussal korrigáljuk és csak kis részét manuálisan. A hibadetektáláshoz egyrészről célzott módon kialakított jelfeldolgozási eljárásokat használunk fel, másrészről a fonetikai hangleírások ismérveit, tehát olyan adatokat, amelyeket a beszédfelismerő nem vett figyelembe. A hibafeltárás során támpontul szolgált a zöngés periódusok utólagos automatikus megjelölése a hullámformában, a hangokra jellemző specifikus intenzitásértékek kiszámítása, valamint a jellemző hanghosszúság figyelembevétele a hangkörnyezet függvényében [6].

A hibakeresés során kialakított algoritmusok a hangra jellemző paramétereket hasonlítják össze a tényleges hullámformával a korábban bejelölt két hanghatár közötti szakaszon. A kategorizálás kétszintű.

5.1. Első szint

Elsőként a durva hibákat szűrjük ki, a rosszul jelölt hangokat keressük. Minden hangot és hanghatár jelölést egymással összevet az erre a célra fejlesztett algoritmus és azokat a hangokat tekintjük hibás jelölésűnek, amelyeknek az akusztikai tartalma nagyrészt (75%) egyáltalán nem felel meg a hang fonetikai leírásának (például egy zöngétlen zárhangnak titulált hangsorrész 80%-ban zöngés periódusokat tartalmaz; egy zörejes, nagy energiájú hangsorrész szünetnek van jelölve). Ez a kereső szoftver a fonetikai jellemzés alapján kategorizál.

Néhány esetben ellentétbe kerülhet az elméleti fonetikai kategorizálás és a hang valóságos fizikai tartalma (ez beszélőtől függ), amit a tények ellenére nem szabad hibának tekinteni. Például az intervokális helyzetű zöngés zárhangokat (*...fejében...*) néhány helyen hibásnak találta a program, azokban a hangkapcsolatokban, ahol a zárhangban nem lehetett periodikus elemeket felfedezni (így ejtette a beszélő). Ez valójában nem hiba. Az algoritmus fejlesztése során az ilyen hibás döntések elkerülésére előírtuk például, hogy VCV hangkörnyezeti esetekben a zöngés zárhangot minden esetben zöngésnek kell tekinteni, függetlenül a hanghullámban mérhető állapotoktól. Egy további ilyen példa, hogy a /h/ hangra engedélyeztük, hogy zöngés és zöngétlen formában is szerepelhet a hangsorban.

A durva hibákat megjelölő szoftver hibalistát generál, amelyben megadja a hiba hangsorbeli helyét, hangkörnyezetét és fajtáját. A listában sokféle hiba szerepel, számuk általában nagy. A gyakorlat azt mutatta, hogy az ilyen durva hibák megszüntetését nem lehet egyetlen javító algoritmussal elvégezni a hibák sokrétűsége miatt, hibatípusonként külön szubrutinokat kellett fejleszteni. A szubrutinok megírásához kategorizáltuk a listából a tipikus eseteket. Tipikus hibának tekintettük, ha ugyanaz a hiba legalább 12-szer megjelent a listában. A továbbiakban manuális, vizuális és auditív vizsgálatot végeztünk a többször előforduló egyforma hibákra (5-7 ugyanolyan hibát vizsgáltunk) annak megállapítására, hogy algoritmizálható-e a hiba kijavítása. Ha egységes képet mutattak a vizsgált esetek, akkor egy-egy egyszerű algoritmust dolgoztunk ki a hiba javításához. Ezt lefuttatva a beszédadatbázison az adott hibák megszűntek és az összesített hibaszám csökkent.

A négy adatbázis vizsgálata során összesen 18 fajta javító szubrutint dolgoztunk ki a durva hibák csökkentésére. Néhány példát mutatunk a *2. táblázatban*. Feltüntettük azt is, hogy a hibásnak talált hangkapcsolatból összesen hány darab van az adatbázisban, ebből látható, hogy a beszédfelismerő csak néhány esetben adott téves felismerést és címkézést. A szürke mezők jelzik azokat a hibákat, amelyek javítására szubrutint dolgoztunk ki. A táblázatból az is kiolvasható, hogy a legtöbb durva hibát az N2 jelű beszélő mondataiban találtuk, aki jellemzően gyorsbeszédű volt.

2. táblázat A feldolgozott beszédadatbázisokban talált hibás hangkapcsolatok előfordulási adatai

Adatbázis	N1		N2		N3		F1	
hibás hangkapcs.	hibás	összes	hibás	összes	hibás	összes	hibás	összes
/s,sz,c,cs/ + /m,n,ny/	8	1340	1	158	1	72	1	18
/amf/		232	22	923		4		2
/cez/			73	1528		61		
/cij/ + V		104	1	116	1	55		4
/l/ + /end_sil/		292	8	642		9	4	11
/gyez/		1	4	707		101		4
/kil/		152	271	6304		72		9

Az adatok azt mutatják, hogy a hibák előfordulása egyrészről szövegfüggő (más témájú szövegben esetleg elő sem fordul az adott hangkapcsolat), másrészről a beszélőtől is függ.

PÉLDA: /a/ + /m/ + /f/ kapcsolat

HIBA: az /m/ hang teljes egészében az /f/ első felére van címkézve, így a nazális hang zöngétlen szakaszra esik *(3. ábra felső része).*

JAVÍTÁSI szubrutin: az /m/ jobb oldali határát az /f/ kezdetére (zöngés-zöngétlen váltási pont) kell elcsúsztatni, a jobb oldali határt pedig az /a/ /m/ hangkapcsolat belsejében kell kijelölni a következők szerint:

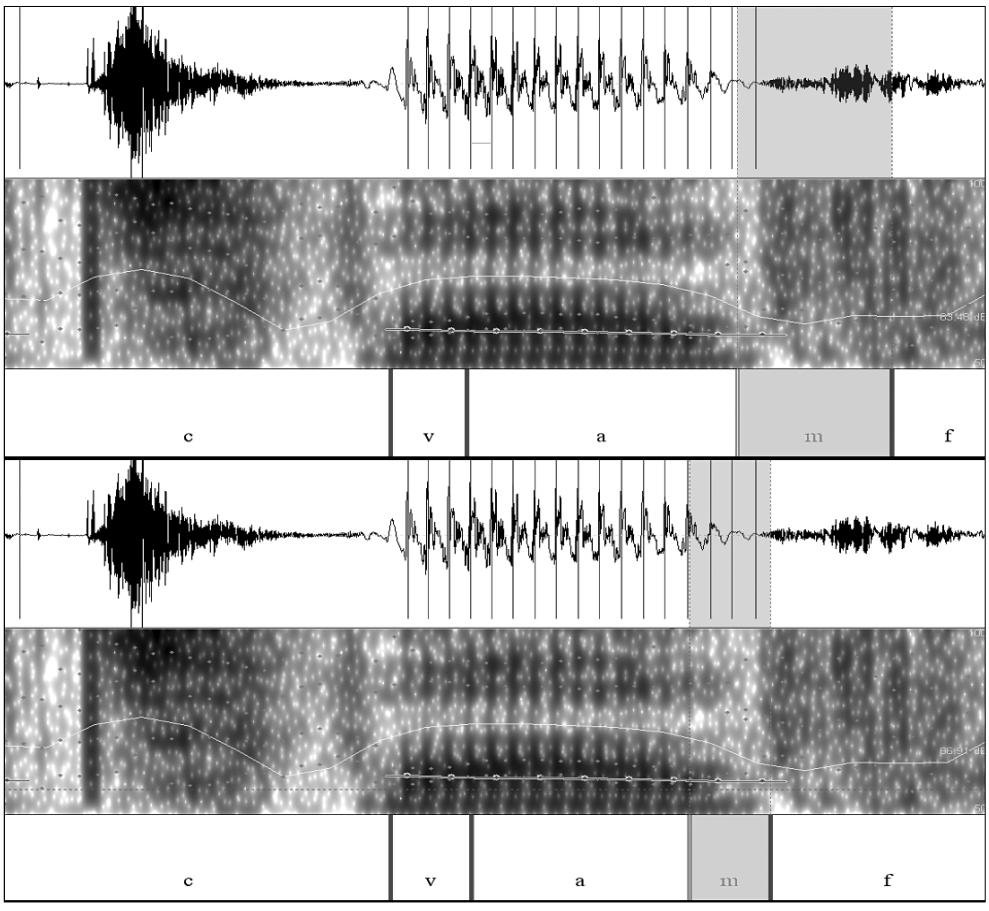
– amennyiben a zöngés szakaszra a /v/ /a/ /m/ hangok vonatkoznak, akkor az időtengely mentén 20-60-20%-ban kell felosztani *(3. ábra lent)* a három zöngés hang teljes időtartamát.
– ha csak /a/ /m/ hangok vannak a zöngés szakaszban, akkor 50-50%-ban kell felosztani.

A durva hibákat kijavító szubrutinokban döntően a zöngés-zöngétlen szakaszok váltakozására támaszkodunk, valamint a hangidőtartamokkal kapcsolatos kutatási eredményekre. Ez utóbbi alapján írjuk elő a több zöngés hangot tartalmazó zöngés szakaszokban az egyes hangokra vonatkozó időarányokat. A durva hibák nagy részét tehát célzott szubrutinokkal megszüntethetjük, a maradékot kézzel kell javítani (ezek a hibák általában egyediek, tehát gépi algoritmust nem éri meg rájuk készíteni). Az ilyen esetekben tehát célszerű felhasználni az emberi közreműködést.

Példák a kézzel javítandó hibákra:

- /s/, /sz/, /c/, /cs/ + /m/, /n/, /ny/ kapcsolatoknál a kapcsolatban résztvevő nazális hang baloldali hanghatára az előző zöngétlen hang végére van címkézve, nagy részben zöngétlen szakaszra. A javítás során a nazális mássalhangzó bal hanghatárát a zöngétlen-zöngés váltópontig kell eltolni, majd a zöngés szakaszra a hanghatárt úgy állítjuk be, hogy a nazális hang és a következő magánhangzó 20-80%-ban osztozzon a kapcsolat időtartamán.
- Felesleges szünetek (nem rövid /sil/, de például zöngés, vagy nagy intenzitású rész /sil/ jelöléssel), ezek általában glottalizáció környékén jelentkeznek például az /o3/ + /sil/ + /o2/ jelölésből /o3/ /o2/ lett a kézi javítás után.
- Rövidre módosult zárszakasz helyes jelölése például a nazális után az /m/ + /b/, /n/ + /c/ esetén
- Átírási hiba (szinkron hiba) a szövegben *kétezer,* a hangban /k/ /e/ /t:/ /o3/ /e/ /z/ /e/ /r/. Ilyenkor több hibásan címkézett hang lesz egymás mellett.
- Számok átírásánál előforduló hiba, a felismerő nincs alternatív kiejtésre tanítva, például: /k/ /i/ /l/ /e/ /n/ /c/ /sz/ /a1/ /sz/, illetve /k/ /i/ /l/ /e/ /n/ /c:/ /a1/ /sz/.

A hibajavítás első szakaszának végére a négy adatbázison összesen 7385 esetben javított a gépi algoritmus durva hanghibát, 4383 esetben pedig a hibás szü-



3. ábra
Példa a nazális hang rossz címkézésére az /amf/ kapcsolatban az N2 jelű adatbázisból (fent). Az automatikus javítás utáni állapotot a lenti részen láthatjuk.

net-jelölések korrekciójára került sor. A kézi javítások száma összesítve 1172. A hibajavítás végére tehát olyan adatbázisaink lettek, amelyekben minden hang a neki megfelelő akusztikai tartalomhoz van jelölve a hullámformában, csak kisebb mértékű hanghatár-elcsúszások lehetnek még jelen, mint hibák. Ezek száma a négy adatbázisban 30658 volt. Ezek javításáról lesz szó a következő pontban.

5.2. A kismértékű hanghatár-elcsúszások feltárása és javítása

A kismértékű hanghatár-elcsúszások feltárásánál és javításánál ugyanazon módszert alkalmaztuk, mint amit a durva hibáknál (lista, tipizálás, szubrutin). Az ilyen hibák nagy része a zöngés-zöngétlen átmenetek határán jelentkezik olyan formában, hogy a tényleges átmeneti ponttól (ami 10-15 ms-os sávra tehető) távolabb van megjelölve a hanghatár, mint ahogy kellene.

Példa:
a *...forintos lebeszélhetőség...* szövegrész szóhatárán az /o/ /s/ hangkapcsolatban az /s/ hang kezdete az /o/ hang 70%-os pontjára van jelölve, tehát zöngés része is van a zöngétlen réshangnak *(4. ábra)*.

A tervezett algoritmussal az összes hasonló hibát sikerült megszüntetni, kézi javításra egyáltalán nem volt szükség.

6. Eredmények

Az új eljárás fejlesztése során világossá vált, hogy a beszédjel gépi címkézéséhez minden olyan információt fel kell használni, ami jellemzi a beszédjelet, tehát kombinálni kell a jelfeldolgozási és a fonetikai eljárásokat, modelleket a pontosabb címkézéshez. Mindezekhez azonban hozzá kell tenni, hogy az itt ismertetett módszer nem alkalmazható automatikus feldolgozásra, mivel emberi tényező is szerepel benne és a feldolgozási idő is hosszúnak tekinthető.

Az új eljárással két-három napi munkával el lehet végezni egy új, több órányi hanganyagot tartalmazó beszédadatbázis címkézését. Az ilyen feldolgozás közel 100%-os pontosságú címkézést eredményez, ami a későbbi használat során fontos tényező lehet.

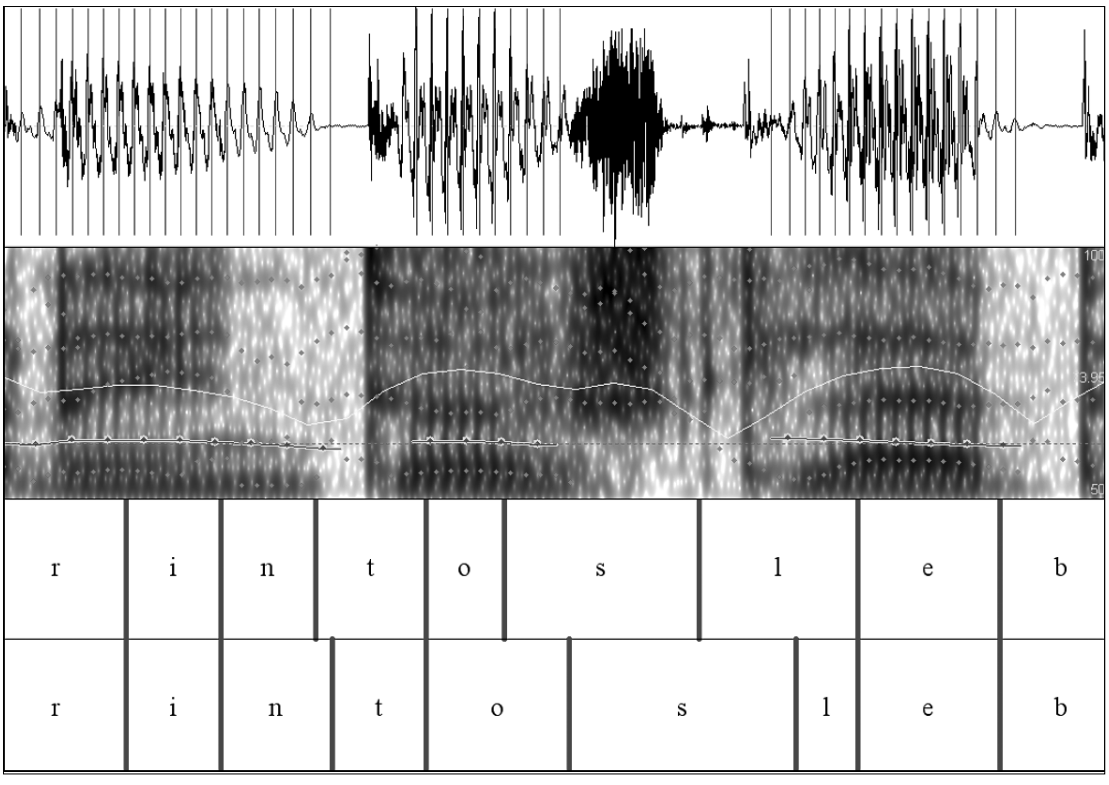
7. Összefoglalás

A több órányi beszédet tartalmazó beszédadatbázisok címkézésénél alapvető gond, hogy a címkézést teljes mértékben emberi erővel nem lehet elvégezni. A cél viszont az, hogy a címkék a lehető legpontosabban legyenek bejelölve a hullámformában.

A fent ismertetett új hibrid eljárást eredményesen lehet alkalmazni az ilyen munkákhoz, szinte hibamentes címkézés érhető el, időigénye is elviselhető (2-3 nap). Az így előkészített beszédadatbázisokban való keresés megbízható eredményeket ad. Ezt fel lehet használni a beszédkutatásban, a beszédszintézisben és a beszédfelismerésben is.

Köszönetnyilvánítás

A kutatást részben az NKTH támogatta a Jedlik Ányos program keretében (TELEAUTO projekt).



4. ábra
Példa a hanghatár elcsúszására (felső hangsor) és annak gépi javítására (alsó hangsor)

A szerzőkről

Olaszy Gábor 1967-ben végzett a BME Villamosmérnöki Kar Híradástechnikai szakán. 1975 óta foglalkozik beszédkutatással, fonetikával (MTA Nyelvtudományi Intézet 2006-ig). Kutatási területei: a beszéd akusztikai szerkezete, szegmentális és szupraszegmentális elemek kutatása, fonetikai modellezés, beszédtervezés, címkézési hibajavító algoritmusok tervezése, többnyelvű szöveg-beszéd átalakító beszédszintetizáló rendszerek tervezése, hullámforma szintézis fonetikai alapjainak kutatása, professzionális beszédkeltők tervezése, készítése, tesztelése. 1983 óta dolgozik a BME Távközlési és Médiainformatikai Tanszék beszédkutató csoportjában is.

Zainkó Csaba 1999-ben végzett a BME Villamosmérnöki és Informatikai Kar Médiainformatika szakirányon és azóta a Távközlési és Médiainformatikai Tanszék Beszédtechnológiai laboratóriumában dialógusrendszerek és az ahhoz kapcsolódó komponensek kutatásával és fejlesztésével foglalkozik. Részt vett az első magyar nyelvű elektronikus levél felolvasó és a számszerinti tudakozó fejlesztésében. Jelenleg a korpusz alapú beszédszintézis technológiájának vizsgálata áll kutatási témájának középpontjában.

Bartalis Mátyás 2005-ben végzett a BME Villamosmérnöki és Informatikai karán, Médiainformatika szakirányon. Oklevele megszerzése óta a BME Távközlési és Médiainformatikai tanszékének Beszédtechnológiai laborjában dolgozik. Fő tevékenysége a beszédszintetizátorokon alapuló alkalmazások fejlesztésében való részvétel, valamint a beszédszintetizátorok adatbázisainak fejlesztése, javítása.

Fék Márk 1997-ben végzett a BME Villamosmérnöki és Informatikai karán, Műszaki Informatika Szakon. 1997-2001 között francia-magyar közös doktori képzésen vett részt a BME-n és a francia ENST-Bretagne-on. Doktori disszertációját beszéd és audio jelek tömörítése témakörében 2006-ban védte meg. 2001-től a BME Távközlési és Médiainformatikai Tanszékén magyar nyelvű beszédszintézissel foglalkozik. Főbb kutatási területei a korpusz alapú beszédszintézis és az érzelemszintézis.

Németh Géza a BME Villamosmérnöki Karán 1983-ban végzett, 1985-ben szakmérnöki diplomát szerzett. 1985-87 között a BEAG Elektroakusztikai Gyárban fejlesztőmérnökként dolgozott, 1987-től a BME Távközlési és Médiainformatikai Tanszékén oktat (Méréstechnika, Kommunikációs rendszerek, Híradástechnika, A jelfeldolgozás elemei, Távközlés, Távközlésmenedzselés, Beszédinformációs rendszerek). Jelenleg a tanszék beszédtechnológiai laboratóriumát is vezeti. Irányító szerepet tölt be a beszédkutatási eredmények gyakorlatba való átültetésében, számos gyakorlati alkalmazást az ő vezetésével fejlesztettek ki.

Mihajlik Péter 1999-ben végzett a BME Villamosmérnöki és Informatikai Karán, Villamosmérnöki Szakon, távközlési főszakirányon és orvosbiológiai technika mellékszakirányon. A gépi beszédfelismeréssel PhD hallgatóként 1999-ben kezdett foglalkozni. 2002 óta a BME Távközlési és Médiainformatika Tanszékén dolgozik – jelenleg tudományos segédmunkatársi minőségben – ahol elsősorban a magyar nyelvű gépi beszédfelismerés témakörében végez kutatásokat.

Irodalom

[1] Paul Boersma P., D. Weenink:
Doing Phonetics by Computer [Comp. software], 2005.
www.praat.org

[2] Mihajlik Péter, Tatai Péter:
Automatikus fonetikus átírás magyar nyelvű beszédfelismeréshez.
In: Gósy Mária (szerk.), Beszédkutatás 2001,
MTA Nyelvtudományi Intézet, Budapest.
pp.172–185.

[3] Mihajlik Péter, Tatai Péter, Gordos Géza:
Automatic Phonetic Transcription and Its Application in Speech Recogniser Training:
A case study for Hungarian
In: Divenyi P., Greenberg S., Meyer G. (ed.),
Dynamics of Speech Production and Perception,
Amsterdam: IOS Press, 2006.
pp.245–262. (NATO Science Series, I.) 374,
Life and Behavioural Sciences.

[4] Olaszy Gábor:
A nazálisok okozta szerkezetváltozás a zár-, rés- és zár-rés hangoknál mássalhangzó kapcsolatokban.
In: Gósy Mária (szerk.), Beszédkutatás 2006,
MTA Nyelvtudományi Intézet, Budapest.
pp.32–43.

[5] Olaszy Gábor:
Mássalhangzó-kapcsolódások a magyar beszédben.
Tinta Kiadó, Budapest, 2007.

[6] Olaszy G., Németh G., Olaszi P., Kiss G., Zainkó Cs., Gordos G.:
Profivox – a Hungarian TTS System for Telecommunications Applications.
International Journal of Speech Technology. Vol. 3-4.
Kluwer Academic Publishers, 2000.
pp.201–215.

[7] Vicsi Klára, Vígh Attila:
Az első magyar nyelvű beszédadatbázis.
In: Gósy Mária (szerk.), Beszédkutatás 1998,
MTA Nyelvtudományi Intézet, Budapest.
pp.163–177.

[8] Vicsi, Klára:
Beszédadatbázisok a gépi beszédfelismerés segítésére,
Híradástechnika 2001/1, Budapest.
pp.5–13.