A BESZÉD SZÁMÍTÓGÉPES FELDOLGOZÁSA

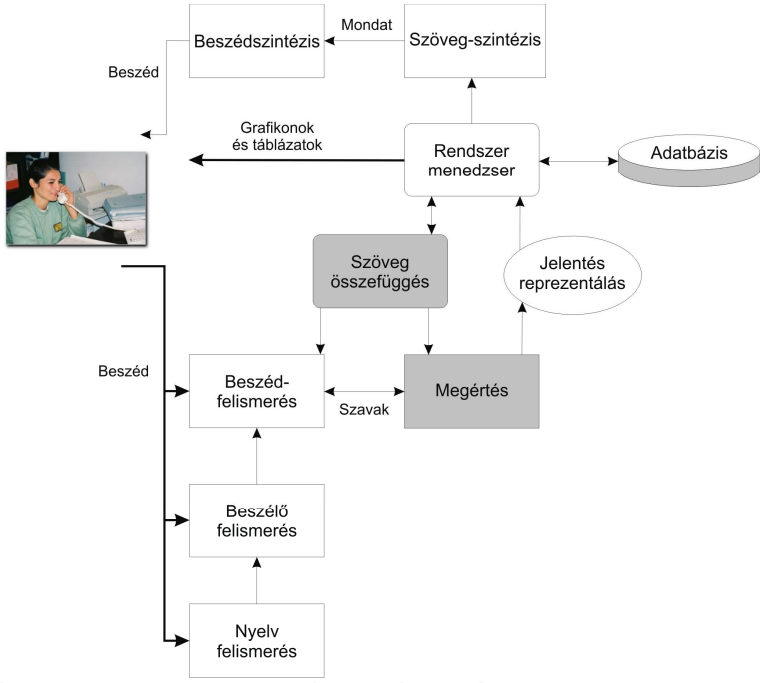
Vicsi Klára, Németh Géza, Szaszák György

1. Bevezetés

Általánosságban a beszédtudomány célja a beszédkommunikáció körfolyamatának komplex leírása, a beszélő gondolatának nyelvi megfogalmazásától kezdve a beszédprodukción át, a létrejött beszéd akusztikai leírásán keresztül, a hallgató beszédpercepciós folyamatain át, a nyelvi tudása alapján a közölni szándékozott gondolat megértéséig.

A beszéd számítógépes feldolgozásánál (beszédtechnológiában) e körfolyamat egyes funkcióit ellátó egységek mesterséges eszközökkel való kiváltása történik. Az egyik fő célja az ember-gép közötti párbeszéd lehetővé tétele. Ezen párbeszéd minél tökéletesebb megvalósításakor nem csak használható, az emberi tevékenységeket támogató eszközök jönnek létre, hanem a megvalósításhoz végzett kutatások, elért eredmények segítenek abban, hogy minél jobban megértsük az emberi beszéd létrehozási és érzékelési eljárásait, az emberi beszéd kommunikációban.

Az ember-gép közötti párbeszéd megteremtése ma többnyire a beszélt nyelvi interfészek megvalósításával történik. Egy beszédinterfész megvalósítása a felhasználó saját nyelvén az ideális, mert ez a legtermészetesebb, rugalmas, hatékony, és gazdaságos formája az emberi kommunikációnak. A beszélt nyelvi interfészek sok különböző technológiát és alkalmazást foglalnak magukban, amint azt a 1.1. ábra szemlélteti. Egy tipikus automatizált beszéd-dialógus rendszer fő komponensei láthatók az ábrán.



1.1. ábra. Egy tipikus automatizált beszéd-dialógus rendszer fő komponensei

Néhány esetben, nem csak az alapul szolgáló nyelvi tartalom produkciója (beszédszintézis) és megértése (beszédfelismerés) az érdekes, hanem a beszélő azonosítása, vagy a beszélt nyelv azonosítása. A beszélő felismerése magába foglalhatja egy specifikus beszélő azonosítását egy ismert populációból, amelynek igazságügyi alkalmazása is lehet, vagy pedig a felhasználó igényelt azonosságának igazolását, ami lehetővé teszi az ellenőrzött csatlakozást helyekhez (pl. számítógépes szoba) és szolgáltatásokhoz (pl. hangos bankszolgáltatás).

Napjainkban az intelligens kommunikációs és információs eszközök (pl. mobiltelefonok, kézi számítógépek, stb.) mérete egyre csökken, míg funkcióik szaporodnak és kezelésük bonyolultabbá válik. A hagyományos eszközök (pl. egér, billentyűzet) kényelmetlenek, vagy a feladat velük meg sem oldható. A beszéddel történő információ csere az egyetlen, ami a kis fizikai méret mellett is megvalósítható megoldásnak tűnik.

Büszkén mondhatjuk, hogy Magyarországon a beszédkutatás mindig a nemzetközi élvonalhoz tartozott. Kempelen Farkas volt az első a világon, aki a 17. század végén, sok évi megfigyelő és kutató munka után megépítette híres beszélőgépét, amellyel beszédhangokat és rövid mondatokat lehetett megszólaltatni. Munkája eredményét az 1791-ben kiadott "*Mechanismus der Menschlichen Sprache*" című könyvében foglalta össze és ezzel Kempelen megalkotta a fonetika és szűkebben a beszédszintézis alapjait. Munkásságát híres magyar beszédkutatók sora folytatta a XIX. és a 20. században. 1916-ban "*Dr. Bánó Miklós okl. mérnök és közgazdasági mérnök Budapesten ...Tetszőleges szöveg reprodukálására alkalmas beszélőgép"* címmel nyújtott be szabadalmi kérelmet, melyet 1919. június 21-én fogadtak el és hoztak nyilvánosságra (Bánó, 1919). Békésy

György a Nobel-díjához vezető kutatásokat a beszédészlelés területén a Posta Kísérleti Intézetben végezte az 1930-as 40-es években.
A 80-as évek elején az MTA Nyelvtudományi Intézete Fonetikai Laboratóriumában megalkották az első magyar szövegfelolvasó szintetizátort (HUNGAROVOX, Kiss, Olaszy, 1984), ami tetszőleges szövegeket tudott a magyar köznyelvi kiejtési szabályoknak megfelelően felolvasni. Az elkövetkező években a beszédtechnológiai kutatás-fejlesztés kiszélesedett Magyarországon és mind a beszédszintézis mind pedig beszédfelismerés területén elsősorban a BME Távközlési és Médiainformatikai Tanszékén (ill. jogelődjeinél) számos elméleti és gyakorlati eredmény született (Németh és tsai, 1998, 2003, 2006, 2007, Tóth, Németh, 2006, 2007, Vicsi, Vig, 1995, 1998, Vicsi et al., 2006; Fegyó et al., 2003, Vicsi-Szaszák, 2005, 2007, Tüske, Z. at al. 2005.).

Az alábbi fejezetekben az ember-gép kommunikáció főbb komponenseit tárgyaljuk, a fő hangsúlyt a beszédszintézisre és a beszédfelismerésre helyezve. A beszédtechnológiára vonatkozó részletesebb magyar nyelvű, interaktív oktató anyag (Olaszy, 2002) több honlapról is ingyenesen letölthető. A www.fonetika.nytud.hu honlapon a magyar beszéddel kapcsolatos interaktív adatbázisok (pl. hangkapcsolatok jellemzői) érhetőek el. A beszédtechnológia részletes áttekintése megtalálható például a Microsoft kutatóinak könyvében (Huang és tsai, 2001). Az emberi tényezők részletes elemzésére jó példa egy amerikai kutatók szerkesztésében megjelent munka (Daryle, Blanchard, 2008).

2. A beszéd szintézise

2.1 Bevezetés

Noha a számítógépes beszédelőállítás technológiája sokat fejlődött, még mindig érvényesnek tekinthetjük azt a szabályt, hogy egy gépi beszédkeltő rendszer szókészletének és minőségének szorzata állandó. Tehát, ha egy kötött témakörre optimalizáljuk a rendszert, akkor jobb minőséget kapunk (sokszor kisebb befektetést is igényel a fejlesztés), mint ha tetszőleges szöveg felolvasására alkalmas megoldást készítünk. Ez a felismerés vezetett a gépi beszédkeltés két alapvető kategóriájának kialakításához (Németh, 2001).

2.2 Kötött szótáras beszéddel válaszoló rendszerek

A beszéd gépi előállítása tekintetében a 90-es évek eleje Magyarországon is forradalmi változásokat hozott. Ez abban nyilvánult meg, hogy nálunk is kezdett kialakulni az ún. kötött szótáras beszélő rendszerek alkalmazása (hangos telefonszámla, hangposta, telefonszám-változás közlése, stb.), melyek alkalmazása a világ legfejlettebb országaiban már a nyolcvanas években megkezdődött. A kötött szótáras rendszerek legegyszerűbb formája, amikor csak előre meghatározott üzeneteket mondatnak ki a géppel. Ezt "tárolt" beszéddel oldják meg. A kívánt közlést egy bemondó felolvassa, ezt digitalizálják, majd visszajátsszák (pl. "*Minden kezelőnk foglalt. Kérjük, várjon!*"). Ez a

technológia jó minőségű beszédet biztosít, de csak addig, ameddig nem kell több tárolt elemet összekapcsolni a kívánt üzenet létrehozásához.

Például egy dátum automatikus felolvasásánál az üzenet tartalma (év, hónap nap, óra perc) változik, vagy egy számla összege, egy kötvény napi árfolyama stb. is mindig más szám kimondatását követeli meg. Ebből következik, hogy bonyolultabb üzeneteket csak több előre eltárolt beszédelem összekapcsolásával lehet összeállítani. Ahhoz, hogy az összefűzött elemek a természeteshez közeli minőséget biztosítsanak, a beszéd frekvencia-, idő-, intenzitás- és dallamszerkezetének folytonosságát biztosítani kell. Tehát például egy számlaegyenleg felolvasó rendszerben a betűkép alapján kézenfekvőnek tűnő 20-30 szó-szintű elem helyett 200-250 építőkockára van szükség ahhoz, hogy a 0-999.999.999 közötti tőszámneveket össze tudjuk rakni (Olaszy, Németh 1999). Ellenkező esetben szaggatott, töredezett kiejtést kapunk. A banki vagy távközlési telefonos információs rendszerekben gyakran hallható ilyen megoldás.

Az emberihez közelítő minőségű kötött szótárú beszédkeltő rendszer létrehozásának a lépései a következők:

- a kimondandó üzenetek témájának áttekintése
- az összefűzendő elemek tárának megtervezése, beleértve az azokat tartalmazó, felolvasandó ún. vivő-mondatok adatbázisát is
- a feladathoz illeszkedő hangú bemondó(k) kiválasztása
- a hangfelvétel elkészítése, majd ebből az akusztikai elembázis összeállítása
- az elemösszefűző algoritmus kidolgozása, programozása és alkalmazásba illesztése
- a rendszer tesztelése és a hangzás végleges beállítása

Ezzel a módszerrel legfeljebb néhány ezer különböző üzenet-elemet tartalmazó rendszert lehet az emberi bemondáshoz közelítő minőségben megvalósítani (pl. dátum- és pénzösszeg felolvasó, kisebb országok menetrend felolvasása, stb.).

2.3. Kötetlen szókészletű szövegfelolvasó rendszerek

A kötetlen szókészlet megnevezés félrevezető lehet, hiszen egy ember is csak az általa ismert nyelve(ke)n, a hozzá közelálló témakörben képes ismeretlen szöveget felolvasni. Pl. egy gépészmérnök kis eséllyel tud orvosi szakszöveget felolvasni és viszont. A közlések stílusa is függ a témakörtől, hiszen máshogy olvasunk fel menetrendet, lakcímet, sport- vagy politikai híreket, verset, mesét, elektronikus levelet, SMS-t, honlap-címet, stb. Ahogy az ember, úgy a jelenlegi gépi megoldások sem képesek teljesen általános megoldást nyújtani, ezért az esetleges adaptációs feladatokat minden témakörhöz meg kell vizsgálni. A gyakorlati alkalmazások további nyelvtechnológiai rendszereket is igényelhetnek (pl. szövegek/szavak nyelvének meghatározása, ékezetek visszaállítása, -Németh és tsai, 1998-). A beszéd érzelmi tartalmának gépi előállítására vonatkozó kutatások magyar nyelven a közelmúltban indultak el (Fék és tsai, 2004, 2005).

A kötetlen szókészletű beszédkeltő rendszerek a következő módon osztályozhatók:

- **Szövegfelolvasó (text-to-speech, TTS):** adott nyelv köznapi szókincsében előforduló szövegek felolvasása (általában kb. 8-10 éves gyermek szókincsének megfelelő)

- **Üzenet felolvasó (concept-to-speech, CTS):** a kifejezni kívánt üzenetre vonatkozó jelekkel ellátott szöveg felolvasása, pl. [Conf_Req] A gépkocsi típusa [Car_Type] Volgswagen Golf?
- **Többnyelvű TTS (multilingual TTS):** azonos építőelemek minél nagyobb halmazának egységes keretben történő felhasználása TTS rendszer megvalósításához több nyelven. Ideális esetben (ami cél és ritkán a valóság) azonos program kód (és hardware) kezeli a különböző nyelvi változatokat, a nyelvfüggő adatok egységes szerkezetű, külső adatbázisban helyezkednek el.
- **Poliglott TTS:** azonos hangon szóló többnyelvű TTS
- **Kötött tematikájú (domain specific) TTS:** csak egy adott témakörű (pl. menetrend, időjárás, szállodafoglalás) szöveg felolvasására alkalmas rendszer. Átmenet egy hagyományos kötött szókészletű és egy általános témakörű TTS rendszer között.

Időnként a beszédszintetizátorok közé sorolják az ún. **képernyő felolvasó (screen reader)** rendszereket is. Ezek azonban csak a számítógép vagy más intelligens eszköz (pl. mobiltelefon) kijelzőjének tartalmát értelmezik vak és gyengénlátó emberek számára. Nem TTS-ek, csak illesztést biztosítanak egy alkalmazói program (pl. szövegszerkesztő) és a TTS között. A kereskedelmi forgalomban a képernyőolvasót és a beszédszintetizátort gyakran együtt árusítják. Ebben a megközelítésben úgy is tekinthetjük, hogy a TTS a szélesebb értelemben vett képernyő-olvasó csomag egy része.

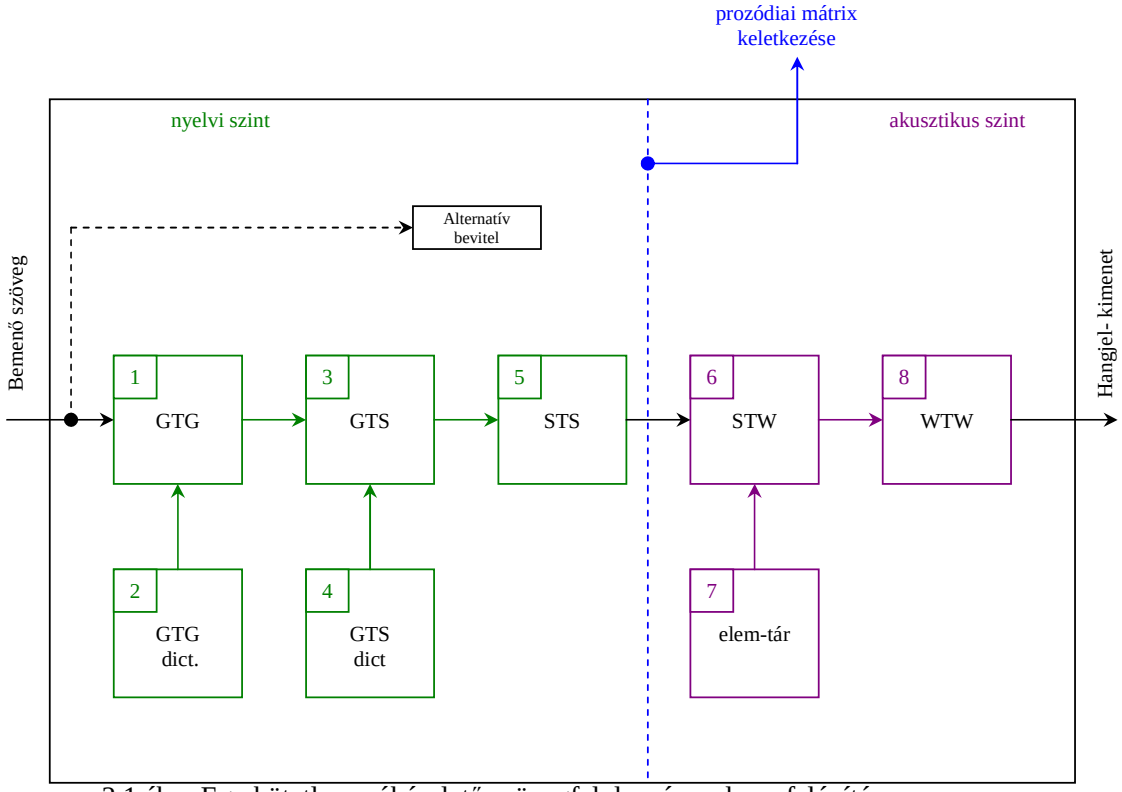
A TTS rendszereket az alábbi szempontok szerint értékelhetjük:

- milyen nyelveken szeretnénk felolvastatni
- milyen szövegeket – egy teljes rendszert általában csak a TTS kimenete alapján ítélnek meg, a bemenetet azonban nem látják
    - néhány lehetséges szövegtípus: általános, szakszöveg, e-levél, SMS, stb.
    - mondattípus: kijelentő, kérdő, felkiáltó, egyéb érzelem kifejezése, segédjelekkel kiegészített, CTS
- milyen minőségben
    - érthetőség (intelligibility)
    - természetesség (naturalness)
    - ezek nem feltétlenül korrelálnak egymással
- milyen hang(ok)on szól a rendszer (pl. férfi-női dramatizált párbeszédek létrehozhatók-e)
- milyen paraméterek állíthatók
    - sebesség
    - hangmagasság
    - suttogás
    - rekedtség
    - szünetek hossza
    - betűzés
- milyen platformokon fusson
    - hardware
    - operációs rendszer (Windows változatok, Linux, Symbian, stb.)
- erőforrásigény, csatornaszám – nem mindegy, hogy mobiltelefonban vagy távközlési szolgáltató központban
- milyen vezérlési felületek, API-k érhetők el

- bővítési, továbbfejlesztési lehetőségek – mit ad hozzá a felhasználó és mit a fejlesztő, pl. speciális rövidítés-feloldó
- milyen speciális igények merülnek fel – pl. visszajelzés egy adott szó kimondásának elején/végén, kimondás állapotának lekérdezhetősége
- milyen támogatást ad a TTS fejlesztő az alkalmazásfejlesztőnek

A 2.1 ábrán egy kötetlen szókészletű szövegfelolvasó rendszer felépítése látható. Az alábbiakban röviden tekintsük át az egyes blokkok működését:

A GTG modul a bemeneti szöveget csak betűket és tagmondat ill. mondathatároló írásjeleket tartalmazó, ún. folyó szöveggé alakítja át. Pl.: „*Az alma123@freemail.hu címre 12:12-kor érkezett üzenet*” szöveget „*Az alma százhuszonhárom, kukac, frímél pont hu címre tizenkettő óra tizenkettő perckor érkezett üzenet.*” alakra hozza. Ez egyfajta egyértelműsítési feladat, ami számos esetben azért is nehéz, mert nincs általános megegyezés a helyes felolvasást illetően (pl. honlap- és elektronikus levél címek, cégnevek esetében, ld. Németh és tsai, 2003). A szöveg belsejében is előfordulhatnak mondatvégi írásjelek, a számok felolvasása szintén bonyolult nyelvi elemzést követel meg. A modul működését segítheti egy szótár (GTG dict), ami rövidítések és speciális kifejezések (pl. @rc kft -> arc káefté) feloldását támogatja. Ez a modul határozza meg az előállítandó prozódia magasabb szintű vezérlését is (pl. egybeolvasandó ún. prozódiai frázisok és azok magas szintű jellemzése, mondat fókusz, stb.).



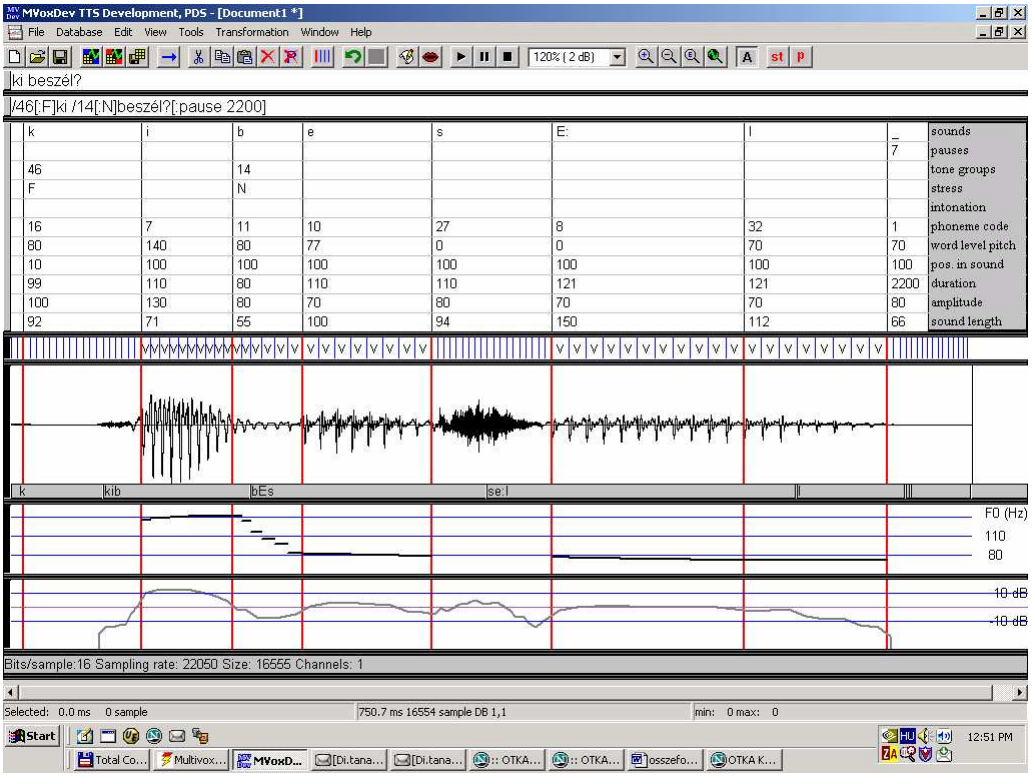
2.1 ábra Egy kötetlen szókészletű szövegfelolvasó rendszer felépítése (Olaszy és tsai, 2000)

1. GTG : Grapheme to grapheme (írásjel→betű)
2. GTG dict. : GTG (kivétel-szótár)
3. GTS : Grapheme to Sound (betű→hang)
4. GTS dict. : GTS szabálylista és szótár
5. STS : Sound to Sound (hang→hang)
6. STW : Sound to Wave (hang→hanghullám)
7. elem-tár : hangelem-tár, akusztikai adatbázis
8. WTW : Wave to Wave (hanghullám-feldolgozás)

Az 1, 3, 5 és 8 modulok elvileg lehetnek nyelvfüggetlenek, viszont a 2,4, 6 és 7 modulok mindenképpen nyelvfüggők. Ha igazán általánossá akarjuk tenni az esetlegesen nyelvfüggetlen modulokat, akkor vagy gépi tanulási algoritmusokat alkalmazunk nagy, előre címkézett adatbázisokon vagy pedig nyelvészeti/fonetikai szabályok alapján működő rendszerek esetén nagyon összetett szabály-leíró nyelv(ek)re van szükség. A beszéd-adatbázisok címkézése is összetett feladat, hibátlan adatbázist még kézi címkézéssel sem lehet létrehozni.

A GTS modul az írott betűk és a kimondandó hangok közötti leképezést végzi el. A beszédszintézisben általában az adott nyelvhez tartozó minimális fonémakészlet helyett a jó minőségű beszédelőállításhoz szükséges, tágabb beszédhang-készletet alkalmaznak (pl. a magyarban külön hangként kezeljük a hosszú és a rövid magánhangzókat, a „h" hang különféle változatait, stb.). A hangok jelölésére általában minden rendszer egyedi konvenciókat használ, ami megnehezíti a rendszerek közötti adatcserét és a többnyelvű megoldások közös fejlesztését. Esetenként köztes megoldásként lehetőség van a SAMPA (ld. ld. Vicsi: „A beszéd akusztikai fonetikai leírása" c. fejezeteket) jelölésrendszerben történő adat export/importra. Noha a magyar az angoltól eltérően ún. fonetikus nyelv, mégis az írott szövegnek nem egy az egyben felelnek meg a kimondott hangok. Ebbe a témakörbe tartoznak egyebek között a hasonulások, azok a hangok, melyeket röviden írunk, de hosszan ejtünk ill. fordítva, a mássalhangzó torlódások és a betűkép helyes értelmezése szó vagy morféma határon (nagyközség, malacság, egészség). Itt történhet meg a prozódia alacsonyabb szintű meghatározása is (pl. szóhangsúlyok). Ezt a modult is támogatja egy formalizált szabály leíró és kivételeket tartalmazó tár (GTS dict).

Az STS modulban kerül sorra a prozódiai vagy egyéb okokból történő hangnyúlások és rövidülések kezelése, beleértve a szünetek megfelelő beállítását is. Első látásra a szünetek kérdése nem túl fontos, azonban ha alaposabban megvizsgáljuk a kérdést, belátható, hogy a szünetek adják meg a folyamatos beszéd megfelelő tagolását. Például, ha nem tartunk szünetet a mondatok között, akkor nagyon nehezen követhetővé válik a minden egyéb szempontból érthető beszéd is.



2.2 ábra A prozódiai mátrix a Profivox fejlesztői rendszerben a „*Ki beszél?*” mondatra

Az STS modul kimenetén áll elő az ún. prozódiai mátrix, ami meghatározza, hogy az adott bemeneti szöveg alapján, milyen vezérlési információk mellett, milyen hangokat, milyen hosszúságban, milyen intenzitással és zöngés hangok esetén milyen alapfrekvenciával kell megszólaltatni. A Profivox fejlesztői rendszerben előálló prozódiai mátrixra látható példa a 2.2. ábrán. A prozódiai mátrix elemei az ábra jobb oldalán szürkével satírozott megnevezésekkel (sounds, pauses, sounds,...) jelölt sorok. A fejlesztői rendszer lehetővé teszi, hogy a legegyszerűbb módon –szövegbeírással (ld. 1. sor), a szöveget szimbolikus vezérlő információkkal kiegészítve (2. sor) vagy magukkal a hangkódokkal (satírozott rész 1. vagy 6. sora) határozzuk meg az előállítani kívánt szöveget. A rendszerrel a hangok prozódiai jellemzői is széles határok között állíthatók, ami megkönnyíti észlelési kísérleti anyagok előállítását is.

Az STW modul a hangelem-tár (vagy más néven akusztikai adatbázis) elemeiből állítja össze a prozódiai mátrixban előírtak alapján a szintetizált hullámforma első változatát. A hangelem-tár minősége alapvetően meghatározza az egész rendszer működését. Természetesen itt is kompromisszumokat kell kötni erőforrás-felhasználás és minőség között. A Multivox rendszer első változatában (Olaszy és tsai, 1992) 255 darab egészen rövid (8-128ms) időszelet formáns-kódolt változata mintegy 1kbyte memóriaterületet igényelt. Az előállított beszéd viszont az erős tömörítés következtében ugyan érthető, de meglehetősen robotos volt. A Profivox rendszer első változatában (Olaszy és tsai, 2000) természetes bemondásban ún. diád elemeket tároltak. A diádok (difón, diphone) hangpárok átmenetet is tartalmazó egységei. Például az *alma* szó diád

elemei: *_a*, *al*, *lm*, *ma*, *a_* (a _ a szünet jele). 1600 diád esetén 22kHz mintavételi frekvencia és 16 bites lineáris kódolás mellett 6.5Mbyte tárigény keletkezett. A természetesség növelése érdekében vezették be a hosszabb elemek tárolását. Ezek közül a diádot követő szint a triád (trifón, triphone). A beszédszintézisben elsősorban a teljes magánhangzót két mássalhangzóhoz kapcsolódva tartalmazó elemeket (ún. CVC kapcsolatok) alkalmazták. A fenti példánál maradva az *alma* szó az *_al* és a *ma_* triádokból valamint az *lm* diádból állítható elő. Egy magyar nyelvű személy- és cégnév felolvasásra optimalizált rendszerben (Németh és tsai, 2006) az elemtár kb. 10.000 elemet tartalmazott és 8kHz mintavételi frekvencia, 16 bites minták mellett 60Mbyte területet igényelt. Ez a megoldás már rövidebb bemondások esetén az emberihez közelítő minőséget eredményezett.

Mivel a fent ismertetett rendszerekben minden elemből csak egy példányt tárolnak, ezért feltétlenül szükség van arra, hogy jelfeldolgozási megoldások segítségével az adott hangrészletet spektrálisan jól leíró jel időtartamát, intenzitását és (zöngés esetben) alapfrekvenciáját a prozódiai mátrixban előírt értékre hozzák. A jelfeldolgozási algoritmusok fejlődése ellenére ez a módosítás még természetes bemondások tárolása és módosítása esetén is jól hallható torzulásokat eredményez. Ennek elkerülésére merült fel az a gondolat, hogy egy-egy bemondótól olyan nagyméretű, akár több órányi hanganyagot tartalmazó adatbázist vegyenek fel, ami (szinte) minden hangot ill. hangkapcsolatot számos változatban tartalmaz és a szintézis során az adott pozícióhoz valamilyen mérték szerint legjobban illeszkedő változat kerül kiválasztásra. Ezt a megközelítést nevezik korpusz-alapú szintézisnek. Ennek a technológiának egy magyar nyelvű időjárás jelentések felolvasására optimalizált változatában (Fék és tsai, 2006) mintegy 10 órányi hanganyag került az elemtárba stúdió minőségben (44.1kHz, 16 bit). Ennek mérete 3.2 Gbyte. Ez a megoldás lehetővé teszi, hogy az időjárás-jelentések felolvasása során nagy valószínűséggel az adott mondatbeli pozícióba jól illeszkedő egész szavak kerülnek kiválasztásra az akusztikai adatbázisból úgy, hogy további jelfeldolgozásra nincs is szükség. A rendszer az eredeti bemondóra megtévesztésig hasonló bemondásokat is képes generálni.

A WTW modul arra szolgál, hogy az adott alkalmazáshoz illeszkedő formátumra hozza az elemtárból kiemelt, összefűzött (esetleg prozódiailag módosított) elemeket. A leggyakoribb az, hogy a mintavételi frekvenciát és az amplitúdó kódolást kell megváltoztatni (pl. az adatbázist 22kHz-cel vették fel, de a telefonos alkalmazáshoz 8kHz mintavételi frekvencia szükséges). De felmerülhetnek összetettebb kódolási igények is (pl. internet-telefonos, ún. VoIP rendszerekben).

A ma elérhető legtöbb beszédszintézis rendszer adott bemenetre mindig pontosan azonos hullámformát generál, ami pl. tudományos kutatások stimulusaként reprodukálható kísérletek megvalósítását teszi lehetővé. Ha azonban gyakorlati alkalmazásokra gondolunk, hosszabb szövegek felolvasásakor kimondottan zavaró, ha mégoly jó minőségben is, de ismétlődően azonos bemondásokat kapunk. A természetes emberi kommunikáció során ugyanis minden megszólalás egyszeri és egyedi, még a *Jó reggelt kívánok!* típusú közlések is. Ennek az emberi tulajdonságnak a modellezésére a közelmúltban indultak kutatások (Németh, Fék, Csapó, 2007).

2.4 A szintetizált beszéd felhasználási lehetőségei a pszicholingvisztikában

A szintetizált beszéd a pszicholingvisztikai kutatások és alkalmazások hasznos segédeszköze lehet számos területen. Kötött szótáras alkalmazásokkal, ahol a stimulus frekvencia-, idő- és intenzitás szerkezete is előre jól megtervezhető és szükség esetén jól kontrollálható, az észlelési alapfolyamatok jól vizsgálhatók(ld. Mády:Beszédpercepció és pszicholingvisztika c. fejezet). Magyar kutatók már a 80-as évek elején sikeresen alkalmazták ezt a technológiát kisgyermekek hallásvizsgálatára, akiket a hagyományos szinuszos vizsgálójellel csak nehézkesen és lassan lehetett mérni (GOH eljárás, Gósy és tsai, 1985). A formánsszerkezet módosítása például ma már grafikus szerkesztői felületen is lehetséges (Bőhm, Németh, 2006).

Parametrikus kódolás (formáns, LPC, stb.) és az elemtár szerkeszthetősége esetén a kötetlen szókészletű szintetizátorok is különösen jól használhatók ilyen célokra. Az emberi bemondás tárolásán alapuló diádos vagy triádos rendszerek jobb érthetősége miatt célszerűbben alkalmazhatók magasabb szintű észlelési kísérletekhez (pl. dallammenetek, szövegtípusok szerinti kísérletek, érzelmek kifejezése). Ez a terület feltételezhetően a közeljövőben jelentős fejlődés előtt áll, mert a szintetizált beszéd elfogadottságának növelése csak akkor lehetséges, ha a gépi felolvasás a dialógus kontextusának megfelelő stílus megvalósítására dinamikusan képes. A rendszerek finomhangolásával lehetőség van kis eltérések (pl. 5Hz-cel kisebb alapfrekvencia) megvalósítására két minta között. Ez a lehetőség jól használható például a kétfülű hallással kapcsolatos kísérletekben. A pszicholingvisztikai kutatások eredményei eddig is jelentősen hozzájárultak a gépi beszédelőállítás fejlődéséhez. Remélhetőleg ez a folyamatos termékeny visszacsatolás a jövőben is folytatódik.

3. A beszéd számítógépes felismerése

A számítógépes beszédfelismerés átfogó beszédfeldolgozási témakör. Legismertebb célja a beszéd nyelvi tartalmának a meghatározása, és ez alapján a tartalom lejegyzése, vagy szóban történő utasítások végrehajtása, vagy a tartalom alapján a géppel való dialógus megszervezése. Azonban a számítógépes beszédfelismerés célja nem csak a beszélt tartalom felismerése, tehát nem csak az érdekes számunkra, hogy mit mondott a beszélő, hanem az is, hogy ki beszél, vagyis a beszélő személy felismerése, azonosítása, továbbá a beszélő hangulatának, egészségi állapotának a felismerése is.

Egy pszicholingvisztikai kézikönyv olvasói számára főleg a beszéd tartalmának számítógépes felismerése a lényeges, ezért e fejezetben a tartalmi felismeréssel foglalkozunk részletesebben, a beszélő személy, valamint a beszéd érzelmi tartalmának felismerését csak érintőlegesen tárgyaljuk.

3.1. Beszédfelismerési feladatok

Amint a bevezetőben már szó volt róla, a beszédfelismerés meglehetősen tág témakör. Szűkebb értelemben a tartalom felismerését értjük alatta, tágabb értelemben azonban alkalmazások egész sora használ egészében vagy komponenseként beszédfelismerőt. A következőkben röviden áttekintjük, milyen feladatokat oldanak meg a beszédfelismerés témakörében. A felsorolás korántsem teljes, részben mert hely hiányában lehetetlen volna valamennyi alkalmazási területet felsorolni, részben pedig mert a témában járatlan olvasó számára is kellő áttekintést adhat az alábbi felsorolás:

- *Beszélőfelismerés* esetén egyik lehetséges célunk a beszélő személy azonosítása (speaker verification), és ezáltal valamely rendszerhez való hozzáférési jogosultság vizsgálata. A másik lehetséges – bár ritkább – felhasználási terület a felhasználó felismerése, kiválasztása egy előre definiált halmazból (speaker identification) (Gordos-Takács, 1983). A beszélőfelismerés történhet szövegfüggő vagy szövegfüggetlen úton (Furui, 1996). Előbbi esetben a beszélő azonosítása meghatározott és a beszélő által előzetesen ismert beszédelemek alapján történik. A módszer nagy hátránya, hogy a beszélőtől felvétel útján előzetesen rögzített bemondás alapján visszaélésre ad lehetőséget, így a szövegfüggetlen módszer tekinthető biztonságosnak: ekkor a beszélő azonosítása egy előzetesen nem ismert, az azonosítás során a helyszínen képernyőn megadott szöveg bemondása alapján valósul meg. Beszélőfelismerésre – természetesen a megfelelő módosításokkal – az alapvetően a beszédfelismerésben is használt megközelítések használhatók (Furui, 1996), ezeket a későbbiekben részletesen ismertetjük, így a beszélőfelismerés technológiájára külön már nem térünk ki.
- A *beszéddetekció* (angolban leggyakrabban VAD, Voice Activity Detection vagy Speech/Non-speech Detection) szinte minden beszédfelismerő alkalmazás elengedhetetlen része (Tucker, 1992), tudnunk kell ugyanis, mikor beszél a felhasználó és mikor nem, hiszen utóbbi esetben felesleges a felismerő rendszernek működnie (sőt, működése sok esetben hibákhoz vezetne). A csendes beszédszünetek jelzésén kívül szükség lehet a környezeti zajok, sőt a zene beszédtől való elkülönítésére is. Tipikusan ilyen problémával találkozhatunk a híranyag adatbázisokban, ahol a kulcsszó alapú keresés vagy automatikus feliratozás előfeltétele lehet a beszéd, a háttérbeszéd és a zene elkülönítése (Vandecatseye et al., 2004).
- A *kulcsszó alapú keresés* akkor lehet hasznos, ha adatbázisokban beszéd alapú keresést szeretnénk megoldani, ekkor a keresőkifejezés nem szövegesen megadott (begépelt) szekvencia, hanem a rögzített kulcsszó.
- Beszéd alapján történő *nyelvfelismerés*re van szükség többnyelvű beszédfelismerő rendszerekben, amelyekben első lépésként a munkanyelv kiválasztását kell automatikusan megoldani.
- Az *érzelmi töltet felismerése* viszonylag fiatal ága a beszédfelismerésnek (Sebe et al., 2005). Egyelőre az érzelmek durvább osztályozása lehet reális célkitűzés, általában 6-8 ún. alapérzelmet szokás elkülöníteni. A számos felhasználási lehetőségen túl a felhasználó érzelmeinek követése sokat segíthet a dialógusok dinamikus felépítésében, a beszélő érzelmeire adekvát gépi válasz kiválasztásában, ily módon az ember-gép kommunikáció teljesebbé tételében.

3.2. A számítógépes beszédfelismerők osztályozása

A beszédfelismerést számos szempont szerint tovább osztályozhatjuk. Egy lehetséges osztályozást mutat be a 3.1 táblázat (Gordos-Takács, 1983).

3.1 táblázat. Számítógépes beszédfelismerés osztályozása

Osztályozási szempont	Artikuláció	Beszélő	Akusztikus környezet	Szótár mérete	Üzemmód
Beszédfelismerő	izolált szavas	beszélőfüggő	csendes	kicsi (<100szó)	parancsmód
	kapcsolt szavas	beszélőfüggetlen	zajos	közepes (100 -	dialógus alapú

típusok				1000 szó)	
	folyamatos	nem adaptált beszélőadaptáció	telefonos	nagy (>10000 szó)	diktáló
				kötetlen	

Az izolált szavas beszédfelismerő szavak felismerésére alkalmas, használatakor a felhasználónak a szavak között szünetet kell tartania. Kapcsolt szavas rendszerben bizonyos szókapcsolatokat a rendszer már felismer, így részlegesen elhagyhatók a szavak közti szünetek, míg a folyamatos beszédfelismerő (Jelinek, 1969) képes kezelni a folyamatos beszédet, így a természetes nyelvhasználathoz a legközelebb áll. Az izolált szavas felismerők egy lehetséges felhasználási területe dialógusokban számjegyek, megerősítő válaszok, stb. felismerése, míg nagy szótáras diktáló rendszerekben a folyamatos felismerés a követelmény.

A beszédfelismerőnek lehet beszélőfüggetlen és beszélőfüggő formája. Beszélőfüggetlen felismerésről akkor beszélünk, amikor a felismerőt használat előtt igen nagyszámú (> 1000) bemondóval előre betanítanak az adott szókészlet, vagy folyamatos szöveg felismerésére. Olyan rendszerek használatosak olyan esetekben, amikor nem lehet tudni előre, hogy ki lesz az aktuális felhasználó.

A beszélőfüggő rendszeréknél maga a felhasználó tanítja be a rendszert a saját hangjára.

A beszédfelismerő lehet beszélőadaptált, ekkor az egyes felhasználóktól származó beszéddel a felismeréshez használt – a későbbiekben részletesen bemutatandó – beszédhang modelleket a rendszer korrigálja, ily módon lényegében az adott személy akusztikai profiljára szabja azokat, melynek révén a felismerés pontosabb lesz (Padmanabhan et al., 1998). Hacsak lehet, érdemes használni a beszélőadaptációt: diktáló rendszerekben gyakorlatilag elengedhetetlen, bizonyos alkalmazásoknál azonban – például nyilvános információ-lekérdező rendszerekben – kivitelezése gazdaságtalan lenne, hiszen a beszélőadaptáció a felhasználó aktív közreműködésével történik, így jelentősebb időráfordítást igényel.

A beszédfelismerők működése szempontjából rendkívül fontos az akusztikai környezet. Csendes környezetben jó jel-zaj viszonyt tudunk biztosítani, ezért a felismerés pontosabb. Zajos környezetben speciális algoritmusokkal szükséges a beszédfelismerő zajtűrését – robusztusságát – javítani (Acero-Stern, 1990; Stern et al., 1992), a felismerés hatékonysága azonban várhatóan így is romlik a csendes környezettel összehasonlítva, ugyanis műszakilag az emberi percepciónál jóval korlátozottabban tudjuk a zajelnyomást megvalósítani. A telefonos környezet sávhatárolt jellege (300-3800 Hz közötti frekvenciatartomány) miatt szintén megkülönböztetett felhasználási terület, a telefonos beszédfelismerés azonban fontos szerephez juthat telefonos információs rendszerekben.

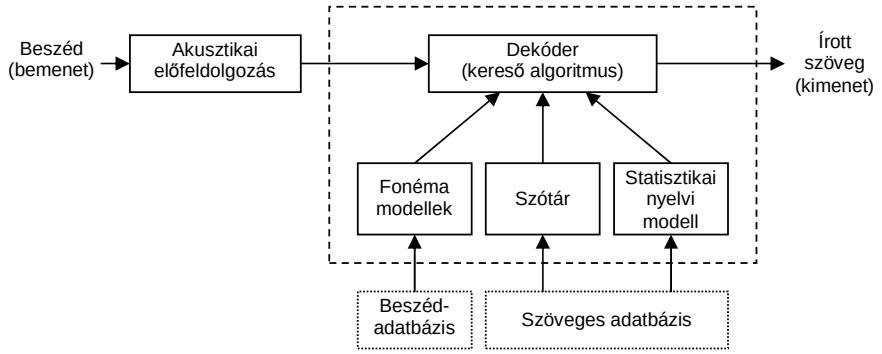
A szótár mérete arányban áll a nyelv modellezési képességével. Nagyobb szótárméret esetén általában hatványozottan bonyolultabb a nyelvi modell, ezáltal a működés is lassul. A szótár méretét jelentősen befolyásolja (korlátozza) egyrészt a valós idejű működés követelménye, másrészt az egyes szavak közötti akusztikai hasonlóság, az akusztikai környezet, a beszélőadaptáció léte vagy nem léte is.

Végezetül, a számítógépes beszédfelismerőket alapvetően kétféle üzemmódban használhatjuk: parancsmódban valamilyen eszköz – számítógép – vezérlése oldható meg beszédinterfészen keresztül, diktáló üzemmódban pedig szövegszerkesztés jellegű munkához kapunk támogatást. Vegyük észre, hogy előbbi tipikusan izolált szavas, utóbbi pedig folyamatos felismerőt igényel. A dialógus alapú üzemmód jóval intelligensebb

rendszert, a beszédet nem csak átalakító, de azt mélységében értelmező és megértő rendszert feltételez, ezért napjainkban ebben a körben csak szerény tudású, kísérleti alkalmazásokkal találkozhatunk.

3.3. A számítógépes beszédfelismerők felépítése és működése

A számítógépes beszédfelismerők alapvető működését követhetjük végig a 3.1 ábrán, mely egy statisztikai alapú beszédfelismerő rendszert (Jelinek, 1976) mutat be. A bemenetre kerülő beszédet akusztikai szintű előfeldolgozásnak vetjük alá, ezután következik a dekódolás, melynek során a fonémák modelljeit, a szótárat és a nyelv szintaktikai viszonyait statisztikai alapokon leíró nyelvi modellt használjuk fel. A fonéma modellek, a szótár és a nyelvi modell egyfajta tudást visz a rendszerbe, melynek számítógépes betanításához beszéd-, illetve szöveges adatbázisokra van szükség. A beszédadatbázisnak a nyelvben előforduló valamennyi fonémát és fonémakapcsolatot tartalmaznia kell statisztikailag megfelelő lefedettséget adva (Roach et al., 1996). Hasonlóképpen, a nyelvi modellnek meg kell felelnie a beszédfelismerő használati területére jellemző szóhasználati szokásoknak.



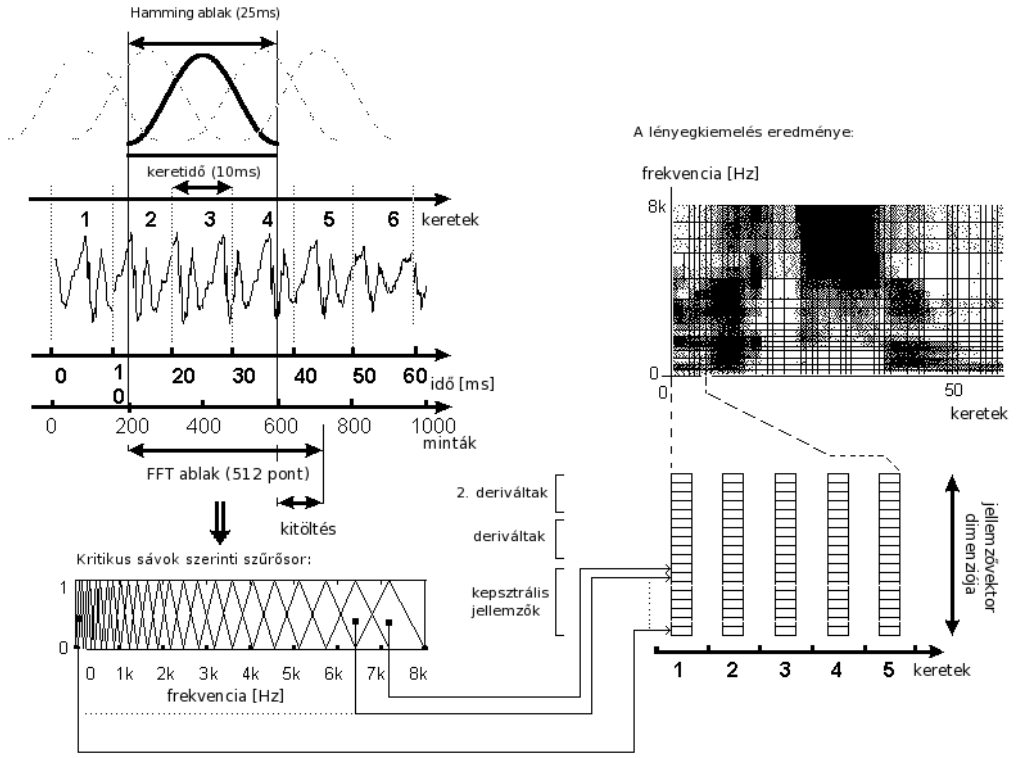
3.1 ábra. A statisztikai alapú beszédfelismerők felépítése és működése

3.3.1. Akusztikai szintű előfeldolgozás a beszédfelismerésben

A számítógépes beszédfelismerőkben az ún. akusztikai modul feladata a beszéd rögzítése, majd a redundáns beszédjelből azon jellemzők kinyerése, amelyek alapján a felismerés a legoptimálisabban elvégezhető. Ez a 2.1 ábrán is látható akusztikai előfeldolgozás, amelyet szokás lényegkiemelésnek is nevezni (vö. Gordos, Takács, 1983). Célunk, hogy olyan, az egyes beszédhangokat jól elkülönítő jellemzőket tartsunk meg, amelyek a beszédet a lehető legtömörebben reprezentálják érdemi információveszteség nélkül. A tömörítés szükségessége az eredendően hatalmas számítási igény kordában tartása miatt is felmerül, a legtöbb alkalmazáshoz kívánatos ugyanis, hogy a beszédfelismerő valós időben is működőképes legyen. Az akusztikai előfeldolgozás egy lehetséges algoritmusát az alábbiakban mutatjuk be (lásd a 2.2 ábrát is).

Első lépésként a beszédet mikrofonnal elektromos jellé alakítjuk, majd a beszédjelet digitalizáljuk, hiszen számítógéppel csak így tudjuk feldolgozni. Ennek során általában

elegendő a 16 KHz-en végzett mintavételezés és a 16 bites kódolás. (Egyes alkalmazásokhoz – különösen, ha a beszédjelet valamilyen távközlő hálózaton keresztül rögzítjük vagy azon akár analóg, akár digitális formátumban továbbítani kívánjuk – ezek az értékek a 8 KHz, 8 bit értékekig is csökkenhetnek.) Következő lépésként a jelet frekvenciatartományba transzformáljuk, erre a Fourier transzformáció digitális jelekre használatos, futási időben optimalizált változatát (FFT) használjuk, így kapjuk a gördülő spektrumot (.ld. „A beszéd akusztikai fonetikai leírása"). A számítások során a 2 valamely hatványával megegyező számú mintát valamilyen ablakfüggvénnyel – leggyakrabban az 2.2 ábrán is látható Hamming ablakkal – súlyozzuk. Tipikusan 20-30 ms hosszúságú, gördülő, azaz az időtengely mentén a beszédjel hosszában végigfutó időablakkal átlagolunk. A folyamatot a 2.2 ábra bal felső sarkában követhetjük végig. (Az eredményként előálló gördülő spektrumot korábban a „A beszéd akusztikai fonetikai leírása"fejezetben láthattuk.) Ezt követően végezzük a tömörítést. Ennek során egyik lehetséges megközelítésben a jelet ugyanolyan elemzésnek vetjük alá, amilyet az emberi fül is végez (ld. Mády: Beszédpercepció és pszicholingvisztika és Vicsi: A beszéd akusztikai fonetikai leírása c. fejezet; Cohen, 1989), azaz kritikus sávok szerinti szűrősoros elemzést végzünk. A 2.2 ábra bal alsó sarkában egy ilyen, a Mel skála szerinti kritikus sávokat reprezentáló Bark szűrősor, jobb felső sarkában a szűrősoron átvezetett spektrum eredő képe látható. Az egyes szűrők kimenetein egy számszerű értéket kapunk, amely a szűrő sávjának megfelelő frekvencia-intervallumba eső energia (háromszög) ablakkal súlyozott összege. A szűrőkimeneteket logaritmizálva és ismételt transzformációnak alávetve (pl. diszkrét koszinusz transzformáció) a *kepsztrum*ot (Bogert et al, 1963) kaphatjuk meg, melynek számszerű értékeit jellemzővektorokba, más néven keretekbe foglaljuk. A jellemzővektorokba általában az egyes szűrőkimenetek első és második deriváltjait is bejegyezzük, illetve a teljes, frekvenciasávokra nem bontott energia értékét és annak első- és másodrendű deriváltjait is hozzávesszük. Egy-egy jellemzővektor kiszámítása között 5-15 ms időnek célszerű eltelnie (keretidő). Az így kapott, jellemzően 30-50 dimenziós jellemzővektorok reprezentálják az emberi beszédet gépi szinten (lásd a 3.2 ábrán jobbra lent). Az eddigiekben bemutatott lényegkiemelési algoritmus „state-of-the-art" technikának is tekinthető, hiszen általa a gépi beszédfelismerésben leginkább használatos és legjobbnak bizonyuló Mel skála szerinti, kritikus sávszélességű kepsztrális együtthatókat nyerjük. A gyakorlatban ezekre az együtthatókra szokás az MFC (Mel Frequency Cepstral) rövidítéssel hivatkozni.



3.2 ábra. Lényegkiemelés és keretképzés beszédjelből. Balra fent: 16 kHz-en mintavételezett beszédjel spektrumának számítása 512 pontos gyors Fourier transzformációval, 25 ms Hamming ablakkal. Balra lent: Mel skála szerinti kritikus sávszélességű szűrősor karakterisztikája a 0-8 kHz tartományban. Jobbra fent: sávszűrt spektrum. Alul: jellemzővektorok (keretek) képzése.

Természetesen az MFC együtthatók számításán kívül léteznek más lényegkiemelési algoritmusok is, ezek közül a lineáris predikció még mindenképp említendő. Lineáris predikciót alkalmazó lényegkiemelésnél (Markel-Gray, 1976; Gordos-Takács, 1983) az emberi hallás helyett a beszéd képzését, azaz a toldalékcső átviteli karakterisztikáját próbáljuk modellezni. Erre használjuk fel a lineáris prediktort, amelynek együtthatói gyakorlatilag leképezik az emberi hangképző szervek „állásait” az egyes beszédhangokra. A lineáris prediktort magát – helyesebben annak számunkra érdekes szintézis szűrőjét – egy csak pólusokkal rendelkező (*all-pole*) szűrőként valósíthatjuk meg, amelynek együtthatói (LPC, Linear Prediction Coefficients) a hangképző szervek állására, ezáltal a kiejtett beszédhang(szakasz)ra is jellemzőek, így a jellemzővektorba kerülnek. A beszéd lényegkiemelt alakja tehát ismét egy vektorsorozat, amely keretidőnként tartalmaz egy-egy újabb elemet. Különbség csak a jellemzővektor tartalmában van az MFC együtthatókhoz képest, amelyek itt a toldalékcső alakját reprezentálják, azaz nem a hallás alapján, hanem a beszédképzésre visszavezetve adják a beszédjel reprezentációját.

A kritikus sávok szerinti feldolgozást és a lineáris prediktorral való beszédkódolást ötvöző módszer is létezik, ez a perceptuális lineáris predikció (PLP, Perceptual Linear Prediction) (bővebben lásd Hermansky, 1990).

3.3.2. Alapvető beszédfelismerési megközelítések

A beszéd-szöveg átalakítás egyik kézenfekvő módja, hogy a felismerni szándékozott beszédelemeket (szavakat) felvétel útján rögzítjük és eltároljuk, gépi felismeréskor pedig ezekhez a referenciamintákhoz hasonlítjuk az elhangzó beszédet. Ehhez általában szükséges, hogy a referenciamintákat és a bemondást időben is megpróbáljuk egymáshoz illeszteni annak érdekében, hogy az artikuláció vagy a beszéd sebességében az egyes személyek közötti, de akár személyen belüli változásokat is kezelni tudjuk, azaz tulajdonképpen lokális zsugorításokat, nyújtásokat kell elvégeznünk mindaddig, amíg el nem érjük a legjobb illeszkedést a referencia és a bemondás között. Az illeszkedés mérőszáma általában spektrálisan értelmezett távolság a minta és a bemondás között. A módszer innen kapta nevét: dinamikus idővetemítés (Dynamic Time Warping, DTW) (Myers-Rabiner, 1981). Előnye az egyszerű megvalósíthatóság, hátránya, hogy leginkább csak szavak, rövid mondatok felismerésére alkalmas, és a szótár mérete – azaz a gép által felismerhető szavak száma – sem lehet több néhány száznál. Napjainkban mobiltelefonokban találkozhatunk ilyen, a felhasználó által bemondott és eltárolt beszédelemeket felismerő alkalmazásokkal.

A puszta dinamikus idővetemítésnél jóval rugalmasabb felépítésű, és nagyságrendekkel nagyobb szótárral rendelkező beszédfelismerés valósítható meg rejtett Markov modellekkel (Hidden Markov Model, HMM), vagy mesterséges neurális hálókkal (Artificial Neural network, ANN). A következőkben ezeket a rendszereket mutatjuk be.

3.3.3. Beszédfelismerés rejtett Markov modellel

A rejtett Markov modellekre épülő rendszerekben (Rabiner, 1989) a beszédfelismerés tisztán statisztikai alapú (Jelinek, 1976), azaz a megfigyelt folyamatból – mely jelen esetben a rövid szakaszokon stacionernek feltételezett beszéd, illetve jellemzően az egyes beszédhangok – mintákat gyűjtünk, majd a rendelkezésre álló minták alapján egy adott eloszláscsaládból olyan függvényparamétereket próbálunk megbecsülni, amelyek jól leírják az egyes beszédhangokat. Egyetlen előzetes ismeretet tételezünk fel, ez pedig a fonetikai értelemben vett abc, azaz, hogy mely beszédhangok vannak jelen a beszédfolyamatban. Semmilyen más nyelvi tudást nem használunk fel, hanem arra alapozunk, hogy ha a modellezést kellően sok minta alapján végezzük, a beszédhangjainkat leíró modellek nagy valószínűséggel pontosak lesznek, azaz lefedik azokat a lehetséges változatokat, amelyeket a tanításhoz használt adatbázisunkba összegyűjtöttünk. Meglehet, hogy ezt a megközelítést a beszédfelismeréssel először találkozó olvasó esetlegesnek érzi, a gyakorlat azonban azt mutatja, hogy napjaink technikai színvonala mellett a rejtett Markov modellekre épülő rendszerek képesek a legjobb beszédfelismerési teljesítményt adni.

Rejtett Markov modellekkel a beszédfelismerési feladat matematikailag az alábbiak szerint fogalmazható meg:

$$Szó_{felismert} = argmax_{minden\ szóra}\{P(szó|X)\},$$

azaz azt a szót (vagy más beszédelemet) keressük, amelyre az X adott akusztikai megfigyelés-sorozat valószínűsége a legnagyobb. Számunkra azonban az X megfigyelés-sorozat ismert, ezért Bayes tétele alapján átalakítva a fenti összefüggést az alábbiak szerint írhatjuk:

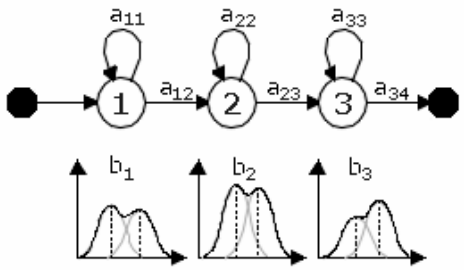
$$Szó_{felismert} = argmax_{minden\ szóra}\{P(X|szó)\ P(szó)\}.$$

Ebben az alakban a P(*X*) tagot a nevezőből elhagytuk. A P(X|*szó*) valószínűséget az akusztikai, a P(*szó*) valószínűséget pedig a nyelvi modell adja meg. Az akusztikai modellnek tehát arról kell informálnia, hogy adott akusztikai megfigyelés az egyes szavakra milyen valószínűségű, a nyelvi modellnek pedig arról, hogy az egyes szavak előfordulásának mekkora a becsült valószínűsége.

3.3.3.1. A beszéd akusztikai-fonetikai modellezése rejtett Markov modellekkel

A lényegkiemeléssel egyfajta mesterséges hallást valósítottunk meg, a következő feladat azon beszédelemek modellezése, amelyeket felismerési egységül választunk. Ezek az egységek lehetnek maguk a szavak, szótagok, leggyakrabban azonban a beszédhangok.

A beszédhangok rejtett Markov modelljei szinte minden esetben három állapotú lineáris struktúrájú (ún. balról-jobbra) modellek (3.3 ábra). A modellezést magát három állapot végzi, valójában azonban két további, szélső állapotot is találunk, amelyek az egyes beszédelem-modellek összefűzését biztosítják. Felismeréskor a rendszer minden keret érkezésekor állapotot változtathat vagy helyben maradhat, bizonyos valószínűséggel. Ezek az ún. állapotátmeneti valószínűségek, melyek becslése a tanítás során történik. Ez a mechanizmus biztosítja az időbeli illesztést a modell és az aktuális keret között. A rendszer az adott (belső) állapotból két keret érkezése között egy megfigyelést bocsát ki, mely tulajdonképpen egy hasonlósági mérték az adott állapotra jellemző jellemzővektor-eloszlás és az aktuálisan érkezett, a külső megfigyelést reprezentáló jellemzővektor között. Lényegében azt mondhatjuk, hogy e hasonlósági mérték a mérőszáma a megfigyelt jellemzővektor és a modellállapot spektrális illeszkedésének. Egy állapotra jellemző jellemzővektor-eloszlást általában sűrűségfüggvényével adunk meg, amelyről feltételezzük, hogy normális (Gauss) eloszlások lineáris kombinációjából áll elő. Ezt szokás kibocsátási valószínűségnek is nevezni.



3.3 ábra. HMM trifón beszédhang modell felépítése és a kibocsátási valószínűségeket megadó függvények szemléltetése. „a” az állapotátmeneti, „b” a kibocsátási valószínűségeket jelöli. Feketével jelöltek az összefűzéshez szükséges állapotok.

Az akusztikai-fonetikai modellek a felismerés során a P(X|*szó*) valószínűségeket szolgáltatják, elkészítésük pedig gépi tanulás eredménye: a rendszer a tanulás során a bemenetére kerülő, ismert tartalmú, fonetikailag átírt beszédet beszédhangonként állapotokra bontja, és az egyes állapotokhoz időben hozzátartozó jellemzővektorokat pedig felhasználja az adott állapot állapotátmeneti és kibocsátási valószínűségeinek a meghatározására (Rabiner, 1989). Az így előálló félkész modelleket aztán iteratívan

finomítja: felismerést hajt végre velük a tanítóanyagon, majd újraszámolja a modellparamétereket. Mindez addig történik, amíg a számított paraméterek értéke szignifikánsan már nem változik tovább. A tanulásnál használt algoritmus a Baum-Welch algoritmus (Baum et al., 1970), amely lehetővé teszi a rejtett Markov modellek paramétereinek maximum likelihood alapú becslését (Bahl et al., 1990), pusztán a rendelkezésre álló megfigyelések – jellemzővektorok – alapján.

Röviden kitérünk arra is, miért éppen három állapotú Markov modelleket használnak a beszédhangok modellezésekor: a válasz a koartikulációs hatás megfelelő lekezelése, amely akár olyan szintű is lehet, hogy egy adott beszédhangra több modellünk is létezik attól függően, hogy az adott beszédhanghoz milyen egyéb beszédhangok kapcsolódnak. Az ilyen ún. trifón beszédhang modellek betanításához jóval több adat szükséges, hiszen a legtöbb nyelvben nagyságrendileg 30-50 egyedi beszédhang modellezése a feladat, míg csak a leggyakoribb beszédhang hármasok lefedése is több ezres trifón elemszámot eredményez. Mindezzel együtt a felismerés szignifikánsan javul, így a módszert elterjedten használják, de a trifónok képzésekor általában klaszterezést alkalmaznak, amelynél legcélszerűbb döntési fák használatával automatikusan meghatározni azon –nem feltétlenül diszjunkt – fonémaosztályokat, amelyeken belül az egyes beszédhangokat már nem különböztetjük meg egymástól. A csoportosítás oka az, hogy valamennyi fonémahármasra elegendő tanítóanyag a beszédadatbázisok méretét hatalmasra növelné, illetve a koartikulációs hatás maga leginkább a szomszédos hangok képzési sajátságaitól függ, amelyek jól osztályozhatóak, s ily módon a gépi klaszterezés alapját képezik.

3.3.3.2. A nyelvi modellezés

A statisztikai felismerésben a nyelvi modell szerepe az egyes szavak, szósorozatok valószínűségeinek becslése, ami a következő összefüggés alapján történhet (N-gram nyelvi modell):

$$P(szósorozat) = P(szó_1, szó_2, ..., szó_N) = P(szó_1)\prod_{i=2}^{N} P(szó_i \mid szó_{i-1}, ..., szó_1).$$

Ehhez az általános esethez képest a gyakorlatban nincs lehetőség arra, hogy szavanként az adott szót megelőző lehetséges szósorozatok mindegyikét számításba vegyük, így általában az adott szót megelőző egy-két szót szokás figyelembe venni, így a bigram, trigram nyelvi modelleket kapjuk. Valójában tehát 2-3 szóból álló szósorozatok előfordulásainak valószínűségét tároljuk, ehhez pedig megfelelően nagy mintahalmazra van szükségünk annak érdekében, hogy a relatív gyakoriságok jó becslést adjanak.

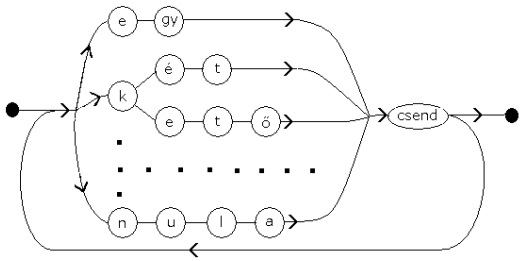
A nyelvi modell tanításához tehát a későbbi felhasználási területet szótárkészletében és szóhasználati stílusában is pontosan lefedő írott korpusz szükséges (hivatkozás Halácsy P. Fejezetére). Lehetőségünk van a korpusz tématerületét behatárolni, s így feladatspecifikus beszédfelismerőt megvalósítani, amely azonban csak az adott nyelvi részhalmazon fog kielégítően működni, igaz ott az általános rendszernél rendszerint jobb teljesítményt is ad.

Általában gondoskodni szükséges arról is, hogy a szótárban vagy a nyelvi modell tanítókorpuszában nem szereplő egyes szavakat is kezelni tudja a rendszer, azaz meg kell őriznünk a modell egyfajta általánosító képességét. Erre használják az ún. simító eljárásokat (részletesebben lásd Bechetti, 1999 vagy Jurafsky, 2000).

A statisztikai nyelvi modellezés kapcsán külön ki kell térnünk a magyar nyelvre is, melynek erős toldalékoló jellege miatt a statisztikai nyelvi modellekkel is csupán a teljes nyelv egy-egy részhalmazát vagyunk képesek lefedni, így ez a fajta megközelítés az ilyen nyelvekre csak igen korlátozottan alkalmazható. Magyar, német és angol nyelvre vonatkozó összehasonlító szövegstatisztikai elemzés szerint egy átlagos szöveg 90%-os lefedéséhez sorra 74.000, 14.500 és 5.000 szóalakra van szükség, ami 97.5%-os fedettség esetén 400.000, 80.000 és 20.100-as értékre nő (Németh, Zainkó, 2002). Tehát egyrészt rendkívül nagy nyelvi szótárat reprezentatívan lefedő anyag összegyűjtése szinte lehetetlen vállalkozás lenne, másrészt a nyelvi modell mérete a szótár bővülésével hatványozottan növekszik. Emiatt veszélybe kerülhet a valós idejű működés, de könnyen elérhetjük a mai számítógépek teljesítőképességének felső határát is, hiszen a nyelvi modellt a felismerés alatt célszerű a számítógép memóriájában tárolni a gyorsabb elérhetőség végett. Tovább bonyolítja a helyzetet a magyar viszonylag kötetlen szórendje is. Alternatívaként a morféma alapú nyelvi modellezés kínálkozik, melynek során azonban nem triviális a szavak belsejében létrejövő hasonulások és a számtalan tőváltozat lekezelése.

3.3.3.3. A beszédfelismerés menete rejtett Markov modelles rendszerekben

Beszédfelismeréskor a nyelvi modell által leírt felismerési struktúra összeállítása történik az első lépésben, azaz felismerési hálózat jön létre. A hálózatban az egyes szavak kapcsolódnak egymáshoz úgy, ahogyan azt a nyelvi modell lehetővé teszi. Maguk a szavak is beszédhangok láncolatából épülnek fel. A 2.4 ábrán egy egyszerű, számjegyek egymásutánját felismerni képes hálózat látható. Az elágazó élekhez az N-gram nyelvi modell súlyokat is rendelhet, ezeket a 3.4 ábrán a jobb áttekinthetőség érdekében elhagytuk.



3.4 ábra. Számjegyek felismerésére alkalmas hálózat

A bemenetre kerülő beszéd keresztülmegy a lényegkiemelésen, majd a rendszer a felismerési hálózat útvonalaira próbálja illeszteni a megfigyelésként érkező jellemzővektor-sorozatot Viterbi algoritmussal (Viterbi, 1967). Az algoritmus a legnagyobb valószínűségű állapotsorozatot keresi meg az adott megfigyelések mellett.

3.4. Alternatívák a statisztikai felismerésre

Már említettük, hogy a rejtett Markov modellek mellett elterjedten használják a beszédfelismerésre a mesterséges neurális hálókat is (Golden, 1996;). Ez utóbbiak beszédhang-modellezési megközelítéseiben nagyobb szerephez jut a fonetikai tudás, mint a rejtett Markov modelles rendszerek esetében, ezzel együtt a mai mesterséges neurális

hálóra alapuló rendszerek is statisztikai elvi alapokon nyugszanak. A mesterséges neurális hálózatok annak ellenére, hogy a biológiai neuronhálókkal vannak analóg tulajdonságai, mégis – sokkal inkább matematikai – számítástechnikai fogalmak. Az elméleti kutatások bebizonyították, hogy a többrétegű neuronhálók univerzális függvény-approximátorok, ennek következtében általános osztályozási feladatokra jól alkalmazhatók. Ami igen fontos tulajdonsága ezeknek a hálózatoknak, hogy párhuzamos feldolgozással működnek.

Természetesen adódik, hogy tulajdonságvektor(-sorozatok) osztályozására is használhatók, megfelelően betanítva felismerik az adott hangrészletet. A mesterséges neurális hálóval történő beszédhang modellezés egy megvalósulásában például az egyes beszédhangok képzésbeli sajátosságai szerint csoportosítunk(**Vicsi, K. – Vig, A. 1995, 1998**). Ily módon – akár hierarchiába is rendezett - fonetikai osztályokat generálunk, és feltételezzük, hogy egy-egy beszédhang képzésekor egy vagy több osztályból képzési jegyeket használunk fel. Ilyen képzési jegy lehet például magánhangzóknál a képzés helye (elől, középen vagy hátul), a nyílt vagy zárt képzésmód, az ajakkerekítés, stb. A mesterséges neurális hálóra épülő akusztikai modul megtanulja, hogyan különítse el az egyes osztályokat egymástól a lényegkiemelt beszédjel alapján, felismeréskor pedig a neurális háló ebbéli tudását használjuk fel. Ahhoz, hogy a háló tanulni tudjon, mintákat is „mutatnunk" kell neki, ez pedig azt jelenti, hogy a tanuláshoz a neurális háló számára is gondoskodnunk kell elegendő mennyiségű, fonetikailag átírt tanító adatról. Beszédfelismeréskor a neurális hálózattal megvalósított felismerő akusztikai moduljának kimenetén a bemenetre kerülő beszédfolyam egyes szakaszain jellemző képzési jegyek jelennek meg. A nyelvi modul feladata az azonosított képzési jegyekből az osztályok rekonstruálása, és a fonémasorozat, szósorozat megadása.(Ide mégis kellene hivatkozás)

A hibrid HMM-ANN beszédfelismerő rendszerek (Cohen et al., 1992) a mesterséges neurális hálók és a Markov modellek előnyeit próbálják együttesen kiaknázni. Ezekben általában a fonetikai osztályozáshoz neurális hálót, a beszédfolyam és a tárolt modell időbeli illesztésére és a nyelvi modellnek megfelelő felismerési struktúra kialakítására pedig Markov láncokat használnak.

3.5. A beszédfelismerés eredményei napjainkban

Kétségtelen, hogy az elmúlt évtizedekben a beszédfelismerés technológiája rendkívüli fejlődésen ment keresztül, ezzel együtt sem szabad azonban elfeledkeznünk arról, hogy a beszédfelismerés korántsem tekinthető megoldott problémának. Jó eredményeket kizárólag jól körülhatárolt alkalmazásokkal sikerült elérni: például a szövegszerkesztőkben használható vagy a számítógép beszéddel történő vezérlését segítő nagy szótáras felismerők amerikai angol nyelvre megközelítik a 100% pontosságot (95-99%) olvasott szövegre, azonban csak akkor, ha a kellően csendes környezetben (megfelelő jel-zaj viszony mellett) használják őket, illetve ha a felhasználó időigényes akusztikai és nyelvi adaptációt hajt végre a felismerés megkezdése előtt. Számjegyek felismerésében a hibásan felismert számjegyek aránya 1% alá csökkenthető beszélőfüggetlen, tehát előzetesen nem adaptált felismerőkben amerikai angol nyelven. Egyre elterjedtebben használják a beszédfelismerőket az egészségügyben leletek elkészítésére is, az emberi ellenőrzés azonban itt sem maradhat el tekintettel a rendkívül magas – gyakorlatilag abszolút – pontosság követelménye miatt. Telefonos beszédfelismerés esetén a kisebb sávszélesség okozta információveszteség miatt a

szótévesztési arányok jelentősen magasabbak. Általánosságban elmondható az is, hogy minél spontánabb a felismerendő beszéd, annál inkább romlik a beszédfelismerés eredménye. Az akusztikai környezet – különösen a külső zajok – szintén rendkívül erősen befolyásolják a felismerés minőségét. A beszédfelismerésben napjainkban mérvadónak tekinthető eredményeket a 3.2 táblázatban foglaltuk össze.

3.2 táblázat. Beszéd felismerésbeli eredmények angol nyelvre, nagy szótárral

Körülmények	Szótévesztési arány (WER)
Olvasott szöveg	<5%
Hírek	~8-10%
Spontán interjúk	~15%
Hétköznapi, spontán beszéd	>30%

Magyar nyelvre a már említett agglutináció problémaköre miatt nagyszótáras beszédfelismerő alkalmazás jelenleg nem létezik. Néhány ezer szavas szótárral, előzetes beszélőadaptáció nélkül leletező – tehát egészségügyi – alkalmazásokban (Vicsi et al., 2005; Fegyó et al., 2003) a helyesen felismert szavak arányát mintegy 95%-ig sikerült emelni.

3.6. A beszédfelismerés és a természetes nyelvek feldolgozásának konvergenciája

A számítógépes beszéd*felismerés* (speech recognition) alapvető célja az elhangzó beszéd lejegyzése írott formában az akusztikai produktum alapján. A mesterséges intelligencia fejlődésével párhuzamosan ez az alapvető célkitűzés egyre inkább kiegészül egyrészt az érzelmek felismerésének igényével, másrészt az automatikus beszéd*értés*sel (speech understanding), amely már elsődlegesen nem a beszéd írott szöveggé alakítására fókuszál, hanem sokkal inkább annak konkrét információtartalmát próbálja megragadni. A beszédértés tehát a beszéddel átadott információ értelmezését is jelenti, beleértve az arra adott valamiféle reakció kiváltását is. A beszédfelismerés és a beszédértés közötti hatalmas különbséget jól érzékelteti a beszédet tanuló gyermek példája, aki a beszédjellel a jelentést állítja szembe (hiszen írni is csak később, az iskolában tanul majd meg). ezzel szemben a számítógépes beszédfelismerő jelenleg „csak” az írásképet (Bechetti-Prina Ricotti, 1999), így a leírt szöveg értelmezése a magasabb – szintaktikai, szemantikai – nyelvi szinteken további, napjainkban összességében még megoldatlannak tekinthető feladat.

A számítógépes beszédértés igényének megjelenésével összhangban a beszédfelismerésbeli feldolgozás a hagyományosan figyelembe vett akusztikai és a ráépülő fonetikai-fonológiai szintről tovább folytatódik a szintaktikai, majd a szemantikai szintek felé (Ainsworth, 1976). Mindez azt is jelenti, hogy az akusztikai szinten újabb paraméterek követése válik szükségessé, így a színképi információ mellett szupraszegmentális paraméterek (alapfrekvencia, energia, temporális viszonyok) is megjelennek a feldolgozásban (Kompe, 1997; Vicsi-Szaszák, 2005), sőt, az akusztikai információ képi információval egészülhet ki, amely az artikuláció képi feldolgozásán túlmenően akár a beszélő személy mimikájának, sőt gesztusainak automatikus követését és értelmezését is magába foglalhatja (Esposito, 2007). E széles körű feldolgozási technológiára utalva szokás a *multimodális beszédfelismerés* kifejezés használata.

A beszédfelismerés és a természetesnyelv-feldolgozás egymást gazdagíthatják. Utóbbi a már említett szintaktikai és szemantikai szintekre való „terjeszkedésben” lehet a beszédfelismerés hasznára, de a beszédfelismerés is hozzájárulhat az eredményesebb természetes nyelvi feldolgozáshoz, például a szövegben kevéssé, de a spontán beszédben annál inkább jelen lévő prozódiai információ kinyerése révén. A közeljövőben várhatóan a két terület egymáshoz jelentősen közeledni fog.

4. Számítógépes és emberi beszédfelismerési modellek és eljárások egymásrahatása

4.1. Bevezetés

Két kutatási terület, amely a beszédfelismerési eljárásokkal foglalkozik, a számítógépes automatikus beszédfelismerés (angolul:automatic speech recognition) és az emberi beszédfelismerés (angolul: human speech recognition). Bár a két kutatási terület szorosan kapcsolódik egymáshoz, céljaik és kutatási megközelítéseik különbözőek.

A számítógépes beszédfelismerésben a központi téma a felismerési hibák számának minimálisra csökkentése. A legtöbb kutatási erőfeszítést az olyan rendszerek fejlesztésére fordítják, amelyek az akusztikus beszédjelekről pontos lexikális átírásokat hoznak létre.

Az emberi beszédfelismerés kutatásában a cél az emberi beszédfeldolgozási folyamat megértése. Ezt olyan elméletek és számítógépes modellek létrehozásával hajtják végre, amelyeket az emberi beszédfelismerési eljárások szimulációjára és magyarázatára lehet használni.

Bár mind a számítógépes beszédfelismerés, mind az emberi beszédfelismerés kutatói a teljes felismerési folyamatot akarják megvizsgálni az akusztikus jeltől a felismert egységekig, egy számítógépes beszédfelismerő rendszer szükségszerűen „elejétől végéig” rendszer, vagyis arra kell képesnek lennie, hogy az akusztikus jelekből a szavakat, mondatokat felismerje, míg az emberi beszédfelismerési modellek legtöbbje az emberi beszédfelismerő eljárásnak csupán részeit írja le. Olyan integrált modell, amely az emberi beszédfelismerési eljárás minden fokozatát lefedi, az utóbbi évekig nem jutott a szerző tudomására. Például a felismerési eljárás egyik alap lépése, amely az akusztikai jeleket átkonvertálja egyfajta diszkrét szimbolikus reprezentációba, többnyire az emberi beszédfelismerési modellekből hiányzik. A beszédpercepciós kutatások számos modellt eredményeztek (ld. Mády:Beszédpercepció és pszicholingvisztika c. fejezet) de egyelőre nem tudtak kielégítően válaszolni arra a kérdésre, hogy hogyan absztrahálja és kapcsolja az agy a beérkező akusztikai jelet magasabb nyelvi egységekké. Következésképp a legtöbb meglévő emberi beszédfelismerés modell nem ismeri fel a valós beszédet. Ez megnehezíti az emberi beszédfelismerés elméleti modelljeinek értékelését a való életbeli tesztelési körülmények között.

A két kutatási területet szeparáltsága ellenére, egyre nő az érdeklődés egymás eredményei iránt. Az emberi beszédfelismerési kutatók remélhetőleg számítógépes beszédfelismerési megközelítéseket fognak beintegrálni egyes parciális modulok helyébe, úgy, hogy komplett „elejétől végéig” modelleket hozzanak létre. A számítógépes beszédfelismerés szemszögéből nézve viszont van remény a felismerési teljesítmény javulására amennyiben az emberi beszédfelismeréssel kapcsolatos egyre gyarapodó alapvető tudás anyagot beépítik a számítógépes beszédfelismerő rendszerekbe.

4.2. A pszichoakusztikai és beszédfelismerési kutatások eredményeinek hatása a számítógépes beszédfeldolgozásra

A gépi hangfeldolgozással foglalkozó szakemberek már eddig is számos esetben figyelembe vették a pszichoakusztikai kutatási eredményeket. Erre igen jó példa a hangtömörítési eljárások között ma a legsikeresebben alkalmazott eljárás, ahol a hangtömörítés a pszichoakusztikai frekvencia és időelfedési kísérletek eredményeire épül (Márki F. 2007). Kialakult az MP3 veszteséges tömörítésen alapuló zenei fájlformátum, kialakult a szabványos MPEG Audio rendszercsalád (ISO/IEC 11172-3-1992), (ISO/IEC 13818-3-1994), ahol igen nagymértékű adattömörítés hajtható végre, anélkül, hogy hallható torzulások keletkeznének a jelben.

Az automatikus beszédfelismerés tématerületén a felismerési algoritmusok kidolgozásánál is, a kutatók több szinten figyelembe vették és sikeresen alkalmazták a pszichoakusztikai és beszédpercepciós kísérleti eredményeket. Például, ahogy már az előző fejezetben láttuk, a modellépítési, osztályozási feladatokat nem a beszédhang nyomás amplitúdó időfüggvényén végzik az automatikus beszédfelismerő rendszerek, mint ahogy ez kézenfekvő lenne, hanem az osztályozást megelőzi még akusztikai szinten, az akusztikai előfeldolgozás. Ezek közül az akusztikus előfeldolgozások közül a legsikeresebbé az a módszer vált, ami modellezi azt az elemzési eljárást, amit mai tudásunk szerint az emberi periférikus hallási rendszer végez. Nagyon jól tudjuk, hogy az emberi hallórendszer érzékeny a jel energiájára, és 'kritikus sávszélességű' felbontásban frekvenciaelemzést végez. Az akusztikai előfeldolgozásban a spektrális elemzéskor a jelet frekvenciatartományba transzformálják, (erre a Fourier transzformáció optimalizált változatát (FFT) használják). Gördülő spektrumot számolnak, amiből megfelelő frekvencia-intervallumba, vagyis a hallási frekvenciaelemzésnek megfelelő 'kritikus sávszélességű' szűrési intervallumba eső teljesítményértékeket, azaz teljesítményspektrumot számolnak (ld. Vicsi: „A beszéd akusztikai fonetikai leírása" c. fejezetet és a jelen fejezetben A beszéd számítógépes felismerése c. alfejezetet.) Ma már a legelterjedtebb automatikus beszédfelismerő rendszerek akusztikai előfeldolgozó, lényegkiemelő modulja e leírt módszer alapjaira épül.

Fonémaszinten, a leghaladóbb számítógépes beszédfelismerő rendszerek a beszédhangokat kontextus-függő fonémaosztályokba sorolják. Fletcher korábbi beszédpercepciós kísérletei jelezték (Fletcher, H. 1953), hogy a hallgatók anyanyelvükön képesek a fonémák felismerésére értelmetlen szótagokban, összefüggéstől függetlenül. Ez annak bizonyítékául vehető, hogy az emberi hallási érzékelés képes kompenzálni a kontextus-függő koartikulációs hatásokat, amelyek az akusztikus beszédadatokban kétségtelenül megnyilvánulnak, és amelyek jelentős problémákat okoztak a spektrális burkoló alapú számítógépes beszédfelismerőkben. Ma a kontextus függő HMM fonémamodellek sikeresen megoldják ezt a fonémaosztályokba sorolást (ld. a jelen fejezetben A beszéd számítógépes felismerése c. alfejezetet).

4.3. A számítógépes beszédfelismerés sikereinek hatása az emberi beszédfelismerési modellekre

A emberi beszédfelismerés szimbolikus elméleteinek (Gaskell, M. G., Marslen-Wilson, W. D., 1997; McClelland, J. L., Elmann, J. L., 1986; Norris, D., 1994) célja, először a bemenő akusztikai jelek leképezése a prelexikális reprezentációba, pl. a

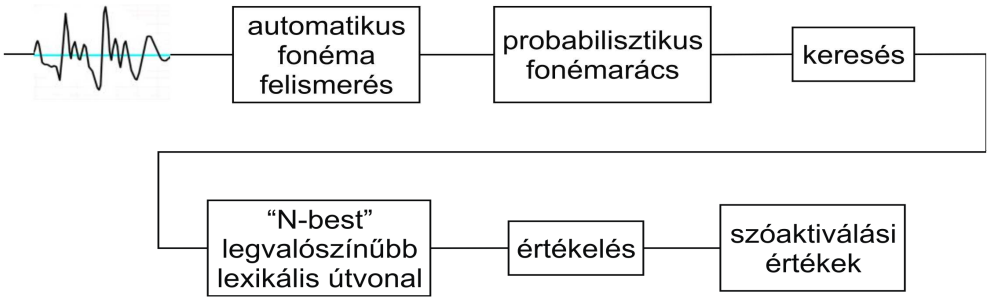
fonémák formáiban, ezután a prelexikális reprezentációk átfordítása lexikális prezentációkba. Így a szimbolikus rendszerekben a beszédfelismerés két szintből áll: a prelexikális szintből és a lexikális szintből. Ebből következően a szimbolikus beszédfelismerési számítási modellek központi követelménye a beszédjel közbenső, szegmentációs reprezentációja. Azonban a legtöbb emberi beszédfelismerés modell híján van annak a modulnak, amely a beszédjelet a szegmentációs reprezentációba konvertálja; ehelyett a bemenetnek egy kézzel készített „hibamentes" diszkrét reprezentációt használnak – abban az értelemben, miszerint a bemenet mindig tökéletesen igazodik a szavak lexikonban lévő szegmentációs reprezentációjához. Így valójában a legtöbb szimbolikus számítási modellben a prelexikális reprezentáció létrehozatalának eljárása csak feltételezett, és fizikai értelemben nincs jelen!

A beszéd bonyolultsága, a beszéddel történő kísérletezés nehézségei, továbbá a beszéd kísérletek költséges volta mind-mind hozzájárulnak a beszéd bemenetű nyelvfeldolgozási modellek eme hiányosságához.(ld. Vicsi: A beszéd akusztikai fonetikai leírása c. fejezet 1. alfejezete). Valószínűleg ehhez az a tény is hozzájárul, hogy a beszélt és írott nyelv feldolgozási folyamatai különösen az alsóbb prelexikális szinteken különböznek egymástól.

Az a tény, hogy a prelexikális eljárás kimenete a beszédjel kézzel készített szegmentációs reprezentációjának formájában érhető el, elhanyagolható is lehetne, ha a beszédjelnek efféle „hibamentes" reprezentációja a valódi beszédfeldolgozásban létezne. A probléma viszont az, hogy az emberi beszédfelismerésben a beszédjel efféle „hibamentes" diszkrét reprezentációja nem hozható létre! Ezt számos kísérlet mutatja (Cucchiarini, C., 1993; Ball, M. J., Rahilly, J., 2002; Shriberg, L. D., et al., 1984). Így pedig a beszédjelek „hibamentes" diszkrét szegmentációs reprezentációja, amelyet a emberi beszédfelismerés legtöbb modellje kíván, nem működik valós beszédbázison.

Röviden, amikor az emberi beszédfelismerés egy integrált számítási modelljét próbáljuk felépíteni, amit meg kell oldani, az az, hogy az integrált modellnek tartalmaznia kell egy valós modult, amely a prelexikális szintet bizonytalanságaival együtt jól szimulálja.

Az emberi beszédfelismerés integrált modellje felé megtett első lépésként 2005-ben kifejlesztették az úgynevezett SpEM (angolul: Speech-based model of human speech recognition) modellt. A SpeM az emberi szófelismerésnek egy „elejétől végéig" modellje (Scharenborg, O., 2005) amely a Shortlist modell (Norris, D., 1994) elméletén alapul, és számítógépes beszédfelismerési technikák használatával építették fel, és a valós beszéd felismerésére alkalmas. (Scharenborg et al, 2005).



4.1. ábra. A SpeM modell grafikus ábrázolása

Az 4.1. ábra mutatja a SpeM modell felépítését. A SpeM három alap modulból áll, amelyek egymás utáni sorrendben működnek. Az első modul egy *automatikus fonémafelismerő*, amely a prelexikális szintet ábrázolja. Itt az akusztikus jelet átalakították a beszédjel szegmentális reprezentációjába, mégpedig a számítógépes beszédfelismerési rendszerekben használatos, és az előző, a „Számítógépes beszédfelismerés" c. fejezetben leírt statisztikai HMM akusztikai-fonetikai modelleket használva (Jelinek, F., 1997). Így, az emberi beszédfelismerés már meglévő legtöbb modelljében használt beszédjel kategorikus lineáris reprezentációjával ellentétben, a SpeM modellben a beszédjelnek egy probabilisztikus reprezentációját hozták létre, egy probabilisztikus fonémagráf formájában. A SpeM *keresési modulja* kiszámítja a bejövő fonémasorozat egyezését a különböző lexikális hipotézisekkel, míg *az értékelő modul* a szóhipotéziseket hasonlítja össze egymással.

A *keresési eljárás* során a legjobb útvonalat határozzák meg, szintén a számítógépes beszédfelismerésben elterjedten használt Viterbi-féle kereséssel a keresési szóterületen át (lásd előző „Számítógépes beszédfelismerés" c. fejezetetet). A keresési terület úgy definiálható, mint egy szófaként felépített probabilisztikus hívásgráf. A folyamatban a fonémarács összes csomópontja feldolgozásra kerül balról jobbra, és minden hipotézist párhuzamosan vesznek figyelembe. A keresési modul által feltételezett szavak egy számmértéket kapnak annak megfelelően, hogy mennyire illeszkednek az aktuális bemenethez. Ha ez a mérték túl nagy a feltételezett szó elvetésre kerül. Akárcsak a számítógépes beszédfelismerő rendszerekben, hasonlóképpen az emberi beszédfelismerésben is csak a legkézenfekvőbb útvonalakat veszik figyelembe. A keresési modul kimenete az SpeM-ben egy N darab legvalószínűbb útvonal, mindegyik egy kapcsolódó útvonalponttal.

Az *értékelő modul* a SpeM-ben ugyan úgy, mint a számítógépes beszédfelismerés statisztikai mintaillesztő technikák legtöbbjében a (ld. A jelen fejezet Számítógépes beszédfelismerés c. alfejezetben) a Bayes-féle valószínűségi szabály alapján számolja ki a szóaktivitást.

Ez a szóaktiválás, viszont, nem alapszik „aktív" gátláson (mint a Shortlist modellben, a gátlás a lexikális reprezentációk között). Ez versenyt modellez a szavak között, de „statikus" módon. Marad a kérdés azonban, hogy vajon ez a statisztikai úton nyert szóaktivitás, vagy pedig az aktív gátlás-e az, ami igazán szükséges. Lehetséges, hogy az előző és a jövőbeli pszicholingvisztikai kísérleti eredmények csak aktív gátlás feltételezésével igazolhatóak. Ha ez történik, a SpeM által elvégzett szóaktiválási számítást át kell alakítani, és felmerül annak kérdése, miszerint hogyan kellene kivitelezni az efféle aktív gátlást.

Mindenesetre a SpeM model nagyon jól rámutat a számítógépes beszédfelismerés és a emberi beszédfelismerés közötti szoros egymásra hatásra. Ráadásul csaknem minden pszichológiai modell feltételezi, hogy az emberi hallgatók párhuzamosan tudják produkálni a szókeresést, ám a meglévő emberi beszédfelismerési modellek általában soros keresést használnak. A SpeM azonban képes a párhuzamos keresésre.

Az emberek képesek szövegösszefüggési információk használatára a beszédfelismerési eljárás során (Zwitserlood, P., 1989). Tehát egy teljes emberi beszédfelismerési rendszernek szimulálni kell a szövegösszefüggési információk használatát is a lexikális szint után. Persze ez a szövegösszefüggési információ nem

korlátozódhat csak a szófrekvenciára, pl. a jelenlegi és az előző szó együttes előfordulásának valószínűségére (Marslen-Wilson, W. D., 1987). A SpeM rendszer csak unigram és bigram nyelvmodellt használ (ld. A jelen fejezet Számítógépes beszédfelismerés c. alfejezetben). Ám, azért, hogy képes legyen modellezni a szövegkörnyezeti információ okozta hatásokat is, ki kell a modellt terjeszteni úgy, hogy a magasabb szintű nyelvmodelleket is képes legyen használni.

4.4. Befejező észrevételek

Ebben az alfejezetben leírt megfigyelések, állítások, bemutatott modellek csak minták arra, hogy megmutassunk néhány olyan egymásra hatást, amivel ez a két tudományterület egymás munkáját már eddig is segíteni tudta. A SpeM modelljének bemutatása azt példázta, hogy egy emberi beszédfelismerési eljárás integrált modelljének kiépítésében hogyan működhetnek közre a számítógépes beszédfelismerés területéről való algoritmusok és technikák.

Mindkét tudományág, a számítógépes beszédfeldolgozás és a kognitív beszédkommunikáció még kibontakozóban van, és nagyon valószínű, hogy a kölcsönös együttműködés mindkét terület számára előnyös lesz. Az a tény, hogy az automatikus számítógépes beszédfelismerés kutatóinak legalább részben sikerei vannak a beszéd nyelvi üzeneteinek számítógépes felismerésében, fel kell, hogy keltse a kognitív tudósok érdeklődését, ugyanakkor a számítógépes beszédfelismeréssel foglalkozó kutatók sem hagyhatják figyelmen kívül a kognitív tudomány ide vonatkozó eredményeit.

Irodalom

Acero, A., Stern, R. M.: Environmental Robustness in Automatic Speech Recognition. Proc. of the ICASSP, Albuquerque, New Mexico, 1990.

Ainsworth, W.: Mechanisms of Speech Recognition. Pergamon Press. Oxford pp 110-124, 1976.

Bahl, L.R,. Jelinek, F., Mercer R.L.: A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Journal of Pattern Analysis and Machine Intelligence (1983); illetve újra megjelent: Readings in Speech Recognition (A. Waibel, K.F. Lee, Eds.)

Ball, M. J., Rahilly, J., „Transcribing disordered speech: The segmental and prosodic layers", Clinical Linguistics & Phonetics, 16, No. 5, 329-34, 2002. Morgan Kaufmann Publishers, San Mateo, CA. pp 308- 319, 1990

Baum, L. E., Petrie T., Soules G., Weiss, N: A maximalization technique occuring in the statistical analysis of probabilistic functions of Markov chains, Ann. Math. Statist., Vol. 41, no. 1, pp. 164--171, 1970.

Bánó, M.: „Tetszőleges szöveg reprodukálására alkalmas beszélőgép", Szabadalmi leírás, no. 74361, Magyar Szabadalmi Hivatal, 1919, június 21.

Bechetti, C., Prina Ricotti, L.: Speech Recognition, Theory and C++ Implemetation. Fondazione Ugo Bordoni, John Wiley. Rome, 1999.

Bogert, B. P., Healy, M. J. R., Tukey, J. W.: The frequency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking, Proceedings of the Symposium on Time Series Analysis (Rosenblatt, M. Ed) Chapter 15, 209-243. New York: Wiley, 1963.

Bőhm, T., Németh, G.: „Algoritmus formánsok követésére, módosítására és szintézisére“, Híradástechnika, 2006/8, 11-16. o.

Cohen, J. R.: Application of an auditory model to speech recognition. Journal of the Acoustical Society of America, 85(6):2623--2629, June 1989.

Cohen, M,. Franco H,. Morgan N,. Rumelhart D,. Abrash V.: Hybrid Neural Network/Hidden Markov Model Continuous Speech Recognition, Proceedings of the International Conference on Spoken Language Processing, Banff, Canada, 1992.

Cucchiarini, C.: „Phonetic transcription: A methodological and empirical study”, Ph.D. thesis, University of Nijmegen, The Netherlands, 1993.

Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro M. (eds): Verbal and Nonverbal Communication Behaviours (LNAI 4775), Springer-Verlag Berlin Heidelberg, 2007.

Fegyó, T., Mihajlik, P., Szarvas, M., Tatai, P., Tatai, G.: Voxenter – Intelligent Voice Enabled Call Center for Hungarian. Proceedings of 8th European Conference on Speech Communication and Technology, Geneva. pp 1905-1908, 2003.

Fék, M., Németh, G., Olaszy, G., Gordos, G.: "Design of a Hungarian Emotional Database for Speech Analysis and Synthesis", Proc. of Affective Dialogue Systems Tutorial and Research Workshop, June 2004., Kloster Irsee, Germany, Springer, ISBN 3-540-22143-3, pp. 113-116.

Fék, M., Szabó, J., Olaszy, G., Németh, G., Gordos G.: Érzelem kifejezése gépi beszéddel, in M. Gósy, M.: Beszédkutatás 2005, Budapest, Október 10-11, 2005. ISSN1218-8727 pp. 134-144

Fék, M., Pesti, P., Németh, G., Zainkó, Cs., Olaszy, G.: “Corpus-Based Unit Selection TTS for Hungarian”, Proc. of Text, Speech and Dialogue, 9th TSD 2006, Sept. 2006, Brno, Czech Republic, Springer, ISSN 0302-9743, pp. 367-373

Fletcher, H.: “Speech and hearing in communication”, The ASA edition, edited by J. B. Allen, Acoust. Soc. Am. 1953

Furui, S.: An overview of speaker recognition technology. In: Automatic Speech and Speaker Recognition (szerk: Lee, C., Soong, F. K. Kuldip, K. P.), Kluwer Academic Publishers, 1996.

Gardner-Bonneau, D., Blanchard H. (Eds.), Human Factors and Interactive Voice Response Systems, 2nd Edition, Springer, 2008. február 19.

Gaskell, M. G., Marslen-Wilson, W. D., „Integrating form and meaning: A distributed model of speech perception”, Language and Cognitive Processes, 12, 613-656, 1997. ISO/IEC 11172-3 International Standard: Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s. Part 3: Audio, 1992, ISO/IEC 13818-3 International Standard: Generic Coding of Moving Pictures and Associated Audio Information. Part 3: Audio, November 1994.

Golden, R. M.: Mathematical Methods for Neural Network Analysis and Design. MIT Press, USA, 1996.

Gordos, G., Takács Gy.: Digitális beszédfeldolgozás. Műszaki Könyvkiadó, Budapest, 1983.

Gósy, M, Olaszy, G., Hirschberg, J., Farkas Zs.: Szintetizált szavak használata a beszédaudiometriában I. és II., Fül-orr-gége gyógyászat 31., Budapest. 1985, 92-96. és 229-233. o.

Hermansky, H.: Perceptual linear predictive (PLP) analysis for speech. Journal of the Acoustical Society of America, 87(4):1738-1752, 1990.

Huang X., Acero, A., Hon, H. W.: Spoken Language Processing, Prentice Hall PTR, 2,001.

Jelinek, F.: A fast sequential decoding algorithm using a stack. IBM journal of Research and Development, 1969.

Jelinek, F.: Continuous Speech Recognition by Statistical Methods. IEEE Proceedings 64:4: 532-556, 1976.

Jelinek, F., "Statistical methods for speech recognition", Cambridge, MA: MIT Press, 1997.

Jurafsky, D., Martin, J. H.: Speech and language Processing. Prentice Hall, USA, 2000.

Kiss, G., Olaszy, G.: „A HUNGAROVOX magyar nyelvű, szótár nélküli valósídejű párbeszédes beszédszintetizáló rendszer, Információ Elektronika, 1984/2, 98-112.o.

Kompe, R.: Prosody in Speech Understanding Systems (LNAI 1307), Springer-Verlag Berlin-Heidelberg, 1997.

Markel, J. D. and. Gray, Jr , A. H.: Linear Prediction of Speech. Springer-Verlag, Berlin, 1976.

Márki, F.: Digitális hangfeldolgozás, Budapesti Műszaki és Gazdaságtudományi Egyetem, Híradástechnikai Tanszék, 2007, hhttp://vibac.hit.bme.hu/documents/58digitalis_hangfeldolgozas_v2.0.pdf

Marlsen-Wilson, W.D.: "Functional parallelism in spoken word recognition", Cognition, 25, 71-102, 1987.

McClelland, J. L., Elman, J. L.: „The TRACE model of speech perception", Cognitive Psychology, 18, 1-86, 1986.

Myers, C. S., Rabiner, L. R.: A comparative study of several dynamic time-warping algorithms for connected word recognition. The Bell System Technical Journal, 60(7):1389-1409, 1981.

Németh, G.: Beszédtechnológia és alkalmazása a távközlésben, in Buzás Ottó szerk., Távközléskultúra, Presscon kiadó, 2001, 200-212. o.

Németh, G., Zainkó, Cs., Bogár, B., Szendrényi, Zs., Olaszi, P., Ferenczi, T.: Elektronikus levélfelolvasó, in Gósy Mária szerk., Beszédkutatás'98, MTA Nyelvtudományi Intézete, 189-203. o. újra nyomtatva Gósy M., Menyhárt K.: Szöveggyűjtemény a fonetika tanulmányozásához, Nikol, 2003, ISBN 963 210 773 X, 231-242. o.

Németh, G., Zainkó, Cs.: Statisztikai szövegelemzés automatikus felolvasáshoz, in Gósy Mária szerk., Beszédkutatás 2000, MTA Nyelvtudományi Intézete, 156-166. o.

Németh, G., Zainkó, Cs., Kiss, G., Fék, M., Olaszy, G., Gordos, G.: "Language Processing for Name and Address Reading in Hungarian", Proc. of IEEE Natural Language Processing and Knowledge Engineering Workshop, Oct. 26-29 2003., Beijing, China, pp. 238-243.

Németh, G.: "Acoustic Company Image and Telecommunications Services", Proc. of Forum Acousticum, 29 Aug.–2 Sep. 2005, Budapest, Hungary, pp. 2633-2637

Németh, G., Zainkó, Cs., Kiss, G., Olaszy, G., Fekete, L., Tóth, D.: "Replacing a Human Agent by an Automatic Reverse Directory Service", Proc. of 15th International Conference on Information System Development (ISD 2006), August 2006, Budapest, Hungary, Springer LNCS, pp. 323-331

Németh, G., Olaszy, G., Bartalis, M., Kiss, G., Zainkó, Cs., Mihajlik, P.: “Speech based Drug Information System for Aged and Visually Impaired Persons”, Proc. of Interspeech 2007, Aug. 2007, Antwerp, Belgium, pp. 2533-2536

Németh, G., Fék, M., Csapó, T. G.: “Increasing Prosodic Variability of Text-To-Speech Synthesizers”, Proc. of Interspeech 2007, Aug. 2007, Antwerp, Belgium, pp. 474-477.

Németh, G., and Zainkó, Cs.: "Multilingual Statistical Text Analysis, Zipf's Law and Hungarian Speech Generation", Acta Linguistica Hungarica, Vol. 49. (3-4), 2002, pp. 385-405.

Norris, D., „Shortlist: A connectionist model of continuous speech recognition”, Cognition, 52, 1989-234, 1994.

Norris, D., McQueen, J. M., Cutler, A.: “Merging information in speech recognition: Feedback is never necessary”, Behavioral and Brain Sciences, 23, 299-3225, 2000.

Olaszy, G.: “Elektronikus beszédelőállítás, a magyar beszéd akusztikája és formánsszintézise”, Budapest, Műszaki Könyvkiadó, 1989, 352 o.

Olaszy, G., Gordos, G., Németh, G.: The MULTIVOX multilingual text-to-speech converter, in: G. Bailly, G., C. Benoit, C. and T. Sawallis, T. (eds.): Talking machines: Theories, Models and Applications, Elsevier, 1992, pp. 385-411.

Olaszy, G., Németh G.,: IVR for Banking and Residential Telephone Subscribers Using Stored Messages Combined with a New Number-to-Speech Synthesis Method, in D. Gardner-Bonneau, D. (Ed.), Human Factors and Interactive Voice Response Systems, Kluwer, 1999, pp. 237-255

Olaszy, G., Németh G., Olaszi, P., Kiss, G., Gordos, G.: "PROFIVOX - A Hungarian Professional TTS System for Telecommunications Applications", International Journal of Speech Technology, Volume 3, Numbers 3/4, December 2000, pp. 201-216.

Olaszy, G. (szerk): Magyar nyelvi beszédtechnológiai alapismeretek, Multimédia CD-ROM, Nikol KKt, 2002

Olaszy, G.: Magyar nyelvi beszédtechnológiai alapismeretek. http://fonetika.nytud.hu/oktat_hu.htm. NIKOL Kkt. 2002.

Padmanadham, M., Bahl, L. R., Nahamoo, D., Picheny, M. A.: Speaker Clustering and Tranformation for Speaker Adaptation in Speech Recognition Systems. IEEE Transactions on Speech and Audio Processing, Vol. 6 / No. 1, January, 1998.

Rabiner, L. R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257-286, 1989.

Roach P. et al.: BABEL: An Eastern European multi-language database. International Conference on Speech and Language Processing. Philadelphia, 1996.

Scharenborg, O., ten Bosch, L., Boves, L., “Early recognition of words in continuous speech”, Proceedings of ASRU, US Virgin Islands, oldalszám 2003.

Scharenborg, O., Norris, D., ten Bosch, L., McQueen, J. M., „How should a speech recognizer work?”, Cognitive Science, Vol. 29, No. 6., pp. 867-918, 2005.

Sebe, N., Cohen, I., Huang, T.S.: Multimodal Emotion Recognition. In: Handbook of Pattern Recognition and Computer Vision, World Scientific, ISBN 981-256-105-6, 2005.

Shriberg, L. D., Kwiatkowski, J., Hoffmann, K., “A procedure for phonetic transcription by consensus”, J. of Speech and Hearing Research, 27, 456-465, 1984.

Formatted: Font color: Auto

Stern, R. M,. Liu, F.-H,. Oshima, Y,. Sullivan T. M, and. Acero, A,: Multiple Approaches to Robust Speech Recognition. Proc. of the Fifth DARPA Speech and Natural Language Workshop, Harriman, New York, February, 1992.

Tóth, B., Németh, G.: "VoxAid 2006: Telephone Communication for Hearing and/or Vocally Impaired People", Proc. of Computers Helping People with Special Needs, 10th ICCHP 2006, July 2006, Linz, Austria, Springer, ISSN 0302-9743, pp. 651-658

Tóth, B., Németh, G.: "Creating XML Based Scalable Multimodal Interfaces for Mobile Devices", Proc. of 16th IST Mobile and Wireless Communication Summit, July 2007, Budapest, Hungary, 5 pages.

Tucker, R.: Voice Activity Detection Using a Periodicity Measure. In: Proc. Inst. Electrical Engineering, vol. 139, no. 4, pp. 377–380, 1992.

Tüske, Z. — Mihajlik, P. — Tobler, Z. — Fegyó., T. Robust voice activity detection based on the entropy of noise-suppressed spectrum. In Proceedings of the Interspeech'2005, Lisbon, Portugal, 2005.

Vandecatseye, A., et al.: The COST278 pan-European. Broadcast News Database. In: Procs. LREC 2004, Lisbon. pp. 873–876, 2004.

Vicsi, K., Víg, A.: Text independent neural network/rule based hybrid, continuous speech recognition, Proc. on EUROSPEECH'95. Madrid p. 201-2204, 1995.

Vicsi, K., Víg, A.: LIAS: Language Independent Automatic Segmentation Technique Using Sampa Labeling of Phonemes, Proc. on First International Conference on Language Resources & Education, Granada Spain, p. 1317-1323, 1998.

Vicsi, K., Szaszák, Gy.: Automatic Segmentation of Continuous Speech on Word Level Based on Supra-segmental Features. International Journal of Speech Technology, Volume 8, Number 4 / December, 2005

Vicsi, K., Velkei Sz., Szaszák, Gy., Borostyán, G., Gordos, G.: Folyamatos középszótáras beszédfelismerő rendszer fejlesztési tapasztalatai: kórházi leletező beszédfelismerő, Híradástechnika 2006/3., (p. 14-20), 2006

Vicsi K., Szaszák Gy., Németh Zs.: Folyamatos magyar beszéd mondatfajtáinak automatikus felismerése, Beszédkutatás 2007, MTA Nyelvtudományi Intézet, Kempelen Farkas Beszédkutató Laboratórium, pp. 162-172, Budapest, 2007.

Viterbi, A. J.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, IEEE Transactions on Information Theory 13(2):260–269, April 1967.

Zwitserlood, P., "The locus of the effects of sentential-semantic context in spoken word processing", Cognition, 32, 25-64, 1989.