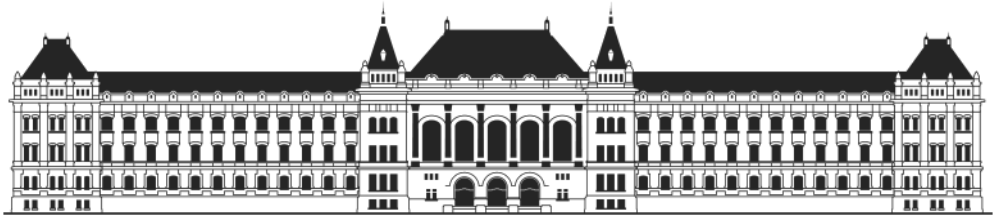
dc_1704_19

Célorientált gépi beszédkeltés interakciós rendszerekben

Az MTA doktora cím
elnyerése érdekében benyújtott értekezés

Németh Géza,
okleveles villamosmérnök, PhD



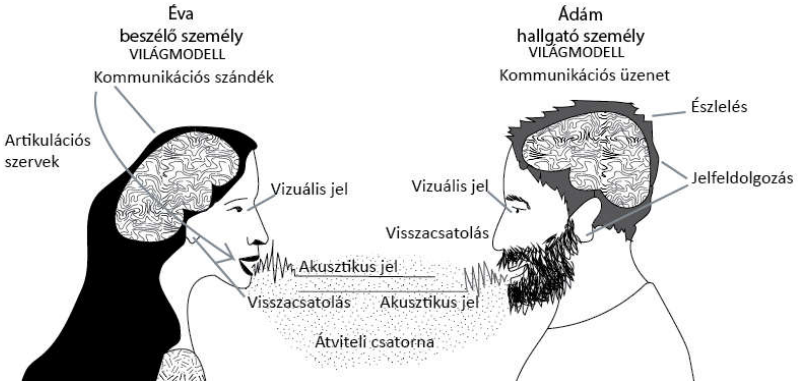
Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék

Budapest, 2019.

Tartalomjegyzék

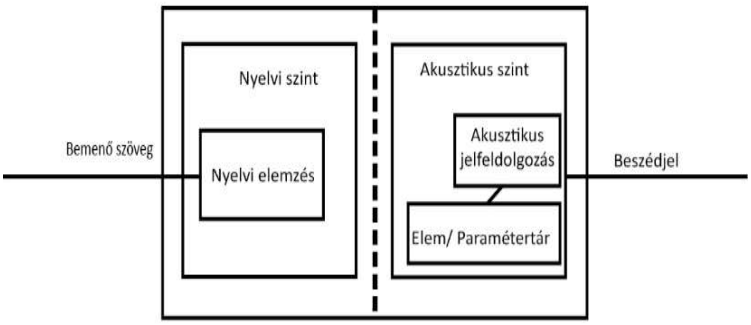
1. Bevezetés

A gépi beszédkeltés a beszédtechnológia tudományterületének egyik ága. Az 1. ábrán láthatjuk a természetes beszédlánc egyszerűsített modelljét. Az emberi kommunikációnak számos alapvető feltétele van. A két partnernek a világról alkotott modellje nagymértékben meg kell egyeznie. Ez a modell hosszú időszak tanulási folyamata révén alakul ki. A modellhez kapcsolódóan fogalmazódik meg az agyban a beszélő személy kommunikációs szándéka, ami a beszédszerveken keresztül alakul fizikai jelekké (elsősorban akusztikus és vizuális formában). Ezek a fizikai jelek egy átviteli csatornán (természetes közegben a levegőn, gépi megoldásnál valamilyen átviteli rendszeren keresztül) jutnak el a hallgatóhoz. A hallgató személy érzékszervei adják tovább a megfelelő biológiai jelfeldolgozás után az észlelés számára az információt. A kommunikációs üzenet értelmezése a hallgató személy világról alkotott modelljéhez kapcsolódóan alakul ki. A beszédkommunikáció alapvető jellemzője, hogy a beszélő és a hallgató szerepe időről időre felcserélődik, így információelméleti szempontból visszacsatolt rendszerről beszélhetünk. Megjegyzendő, hogy az egészséges beszélő személy saját maga is hallja a beszédét és ennek is fontos szabályozó szerepe van (pl. a hangerő meghatározásban). A továbbiakban az akusztikus csatorna szerepével foglalkozunk, mert a gépi feldolgozásban általában annak van elsődleges szerepe.



1. ábra. A természetes beszédlánc egyszerűsített modellje

Beszédtechnológiának a természetes beszédlánc egy vagy több elemének gépi megvalósítását tekintjük [1]. A beszédtechnológia interdiszciplináris tudomány, számos bölcsészeti (pl. nyelvtudomány, fonetika, pszichológia), természettudományi (pl. fizika, matematika) és műszaki területet (pl. akusztika, jelfeldolgozás) érint.



2. ábra. A gépi szövegfelolvasás általánosított modellje

A jelen disszertációban a beszédkeltés gépi modellezése tématerületén a PhD fokozat megszerzése óta elért tudományos eredményeimet foglalom össze. Az elért eredmények emberi közreműködéssel, úgynevezett meghallgatásos tesztekkel értékelhetők, objektív értékelések (küszöb, intervallum stb.) a generált beszéd minőségének megállapítására csak részlegesen alkalmazhatók.

A gépi szövegfelolvasás (Text-To-Speech, TTS) általánosított modellje a 2. ábrán látható. A nyelvi szinten a bemenetre kerülő szövegből meghatározzuk a kimondandó hangokat és azok alapvető prozódiai jellemzőit (időtartam, intenzitás, zöngés hangok alapfrekvencia menete). Az akusztikai szinten pedig a rendelkezésre álló technológiától függő modellek, az aktuális elemtár és az aktuális jelfeldolgozási algoritmus segítségével (vagy anélkül) előállítjuk a kimeneti gépi beszédjelet.

Az 1980-as évek közepéig a megoldások a hangképző szervek (tüdő, légcső, gége, garat, száj- és orrüreg, ajkak) és az artikulációs folyamat működésének leírásán alapultak [2], [3], [4]. A hangképzés artikulációs (forrás-szűrő) modellezése sikerre vezetett, hiszen a modellel az emberi beszédhez megtévesztésig hasonló hangjelenséget is sikerült létrehozni [5], azonban ezzel a megoldással a fő célt, az automatizált gépi szövegfelolvasás emberre emlékeztető szintjét nem sikerült elérni.

Ezért az 1990-es évek elejétől előtérbe kerültek az emberi beszédképzés eredményeként előálló hullámforma tárolásán, feldolgozásán, módosításán és visszajátszásán alapuló megoldások [6], [7]. Ehhez hozzájárult a számítástechnika fejlődése is. Az ilyen megoldásokkal már olyan gépi felolvasó rendszereket lehetett létrehozni, amelyekkel hosszabb szövegek felolvasása is elfogadható hangminőséggel valósult meg, bár a robotos jelleget még magán viselte (pl. e-levél felolvasás és képernyő felolvasás látássérült emberek számára) [8]. További kutatásaink eredményeképpen szűk tématerületen (pl. időjárás jelentés, menetrend-felolvasás) létrehoztunk az emberi felolvasás minőségét és jellemzőit megközelítő rendszereket [9]. Az elmúlt évtizedben

pedig a forrás-szűrő modell és a hullámforma-alapú megközelítés előnyeinek kombinációját ígérő statisztikai parametrikus beszédszintézis (elsősorban Hidden Markov-Model, HMM és Deep Neural Networks, DNN) kialakulásának lehettünk tanúi [10], [11] és részesei [12], [13], stb.

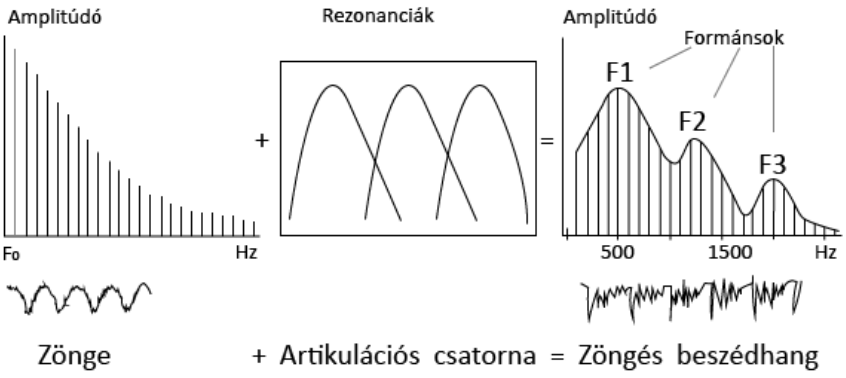
Az is kezd körvonalazódni a kutatások tapasztalatai alapján, hogy az alkalmazási területtől, az ember-gép kapcsolat megoldásától, a felhasználói elvárásoktól függően változhat a géppel előállított beszéd minőségi követelménye. Például egy beszélő robot (bábu, guruló robot) esetén az érthetőség a legfontosabb és kimondottan előnyös lehet, ha nem tökéletesen emberi jellegű, hanem robotos hangzású az előállított hang. A robotikából jól ismert a rejtélyes völgy (uncanny valley, [14]) hatás, mely szerint az emberre hasonlító gép egy bizonyos hasonlósági fokig pozitív érzelmi hatást vált ki, de ezután elérhet egy letörési pontot, ahol már inkább elutasítást okoz az emberben (zombinak tekintjük). Éppen ezért a tökéletes gépi beszéd létrehozásához és annak elfogadásához nemcsak a beszédkeltés mechanizmusát, hanem az agy működését szemantikai szinten is meg kell(ene) értenünk. Ameddig nem érünk el erre a szintre, addig az éppen aktuális felhasználást figyelembe véve és az a priori rendelkezésre álló információk alapján célszerű a feladathoz illeszteni a gépi beszédkeltés megfelelő változatát. Így lehet optimálisabb ember-gép interfészt megvalósítani. A jelen dolgozatban egyrészről a PhD fokozat megszerzése óta a jó minőségű gépi szövegfelolvasás három különböző megközelítésen alapuló technológiájával kapcsolatos új kutatási eredményeimet ismertetem. Fontos megjegyezni, hogy az egyes technológiák nem inkrementális jellegű fejlődés eredményeként, hanem a hardver és szoftver fejlődése által lehetővé tett, elvi megközelítésükben jelentősen különböző kutatások eredményeként jöttek léte. Másrészről bemutatom az eredmények felhasználását hatékony ember-gép interfész megoldásokban, valamint műszaki alkotásokban és alkalmazásokban. A tézisekhez kapcsolódó kutatások (társ)témavezetésemmel megvédett PhD disszertációkat is eredményeztek [15], [16], [17], [18] és [19].

Az értekezés 2. fejezetében történelmi áttekintés keretében ismertetem a gépi beszédkeltés különböző megközelítéseit. A 3. fejezetben kutatási célkitűzéseimet foglalom össze. A 4. fejezetben a kutatás eszközeit és módszereit tekintem át. Az 5.-8. fejezetben kutatási eredményeimet foglalom össze téziscsoportonként. Az alfejezetek elején fogalmazom meg téziseimet. A 9. fejezetben a korábban ismertetett tézisek gyakorlati alkalmazásokban és műszaki alkotásokban megtestesülő felhasználását mutatom be. A 10. fejezet egységes szerkezetben foglalja össze téziseimet. Az értekezést köszönetnyilvánítás és irodalomjegyzék zárja. Ennek a bevezetésnek és a következő történelmi áttekintésnek bővített változatát [20] tartalmazza.

2. A gépi beszédkeltés különböző megközelítései, történelmi áttekintés

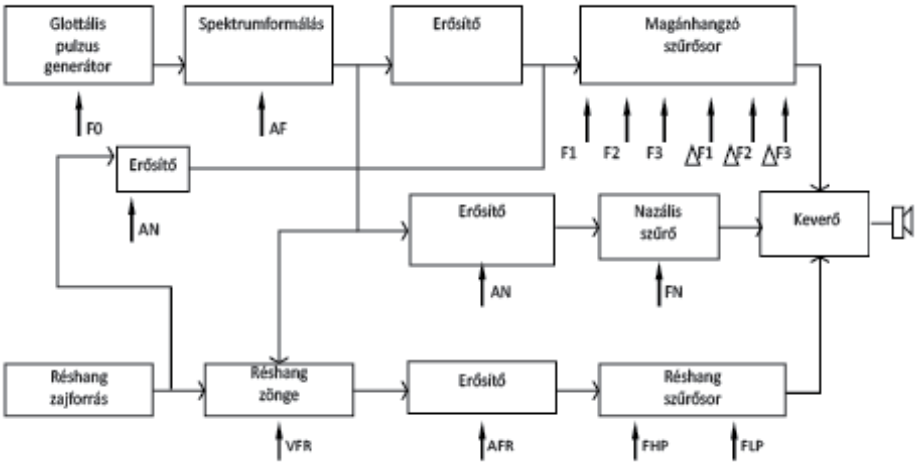
A gépi beszéd-előállítás tudományos alapjait Kempelen Farkas 1791-ben megjelent könyve fektette le. Ennek magyar fordítása 1989-ben jelent meg. [21]. Az első elektromechanikus beszélőgép elvi módszerét is magyar ember találta fel [22]. Nagy média nyilvánosságot kapott a Bell Laboratóriumban az 1930-as években fejlesztett elektromechanikus VODER rendszer [23]. A számítógépes gépi beszédkeltés első megoldásai az 1950-es években születtek meg [24]. A mini- és mikroszámítógépek megjelenésével a hazai kutatók is követhették a nemzetközi trendeket [25], [26], [1].

A különböző elvi megközelítések különböző beszédminőséget és gyakorlati alkalmazási lehetőségeket eredményeztek. Az artikulációs (forrás-szűrő) [27] megközelítés elsősorban az emberi beszédkeltés mechanizmusainak modellezésére volt alkalmas. A formáns-alapú beszédszintézissel (ld. 3. ábra) sikerült kötetlen szókészletű, jól érthető, kereskedelmi forgalmazásra alkalmas, de egyértelműen gépies hangzású, gépi beszédet előállítani.



3. ábra. A gépi beszédkeltés formáns modelljének alapelve

A modell lényege az ún. forrás-szűrő megközelítés (forrás=hangképzés, szűrő=artikuláció). A modellben a zöngés hangokat azonos alapfrekvenciájú (F0) periodikus gerjesztéssel, a zöngétleneket fehérzaj-szerű forrás jellel, az artikulációs csatornát szűrősorral modellezzük. Az így kapott kimeneti jel hullámformája és frekvencia spektruma (főleg a formáns értékek tekintetében, melyek meghatározóak a magánhangzók észlelésében) jó közelítéssel megegyezik a természetes beszédével. A 4. ábrán egy formáns modell részletes blokkdiagramját láthatjuk.



4. ábra. Formánsszintetizátor blokkdiagramja
[28] *alapján*

A formáns-alapú beszédszintézissel sikerült kötetlen szókészletű, jól érthető, kereskedelmi forgalmazásra is alkalmas, de egyértelműen gépies hangzású, szintetizált beszédet előállítani [29]. Ilyen rendszert használt Stephen Hawking, az ismert fizikus egészen haláláig, mivel beszélni nem volt képes. A sok évtizedes használat azt eredményezte, hogy az ő személyét a gép hangkarakterével azonosítják a világban mind a mai napig.

Az artikulációs modellezés korlátjainak kiküszöbölésére indult meg – a számítógépek memóriájának bővülésével és a processzorok gyorsulásával egyidejűleg – a természetes beszéd hullámformájából kiinduló megoldások kutatása [6]. A diád (kiejtett beszédből kivágott két egymás utáni fél beszédhangnyi hullámforma egység) és triád (fél+egész+fél beszédhangnyi egység) elemek összefűzésén alapuló rendszerek hangkapcsolat szintű hullámformákat fűznek össze, majd az így összeállított hullámformán prozódiai módosításokat végeznek jelfeldolgozással, hogy a beszédnek dallama, ritmusa és esetleg hangsúlyozása is legyen [8]. Ezzel a megoldással egyrészt az eredeti emberi hangszínezetre emlékeztető gépi beszédet lehet létrehozni, másrészt viszonylag kis számítási kapacitás mellett lehet változtatható hangkaraktereket kialakítani (férfi, nő). A módszer lehetőséget ad az előállított beszéd sebességének változtatására is. Ennek különös fontossága van a látássérült emberek kommunikációjának szempontjából. Téziseimnek ez a módszer adja az első csoportját.

Újabb módszer – és máig az emberhez leginkább hasonló felolvasást biztosítja – az ún. korpusz-alapú szövegfelolvasó technológia, amely a diád, triád elv továbbfejlesztésének is tekinthető, hiszen szavak, mondatrészek hullámformájának összefűzésével alakítja ki a kívánt beszédjelet. Ennél a módszernél nagy beszédadatbázisra van szükség. Olyanra, amely lefedi azt a témakört, amelyben a gépi beszéd-előállítást használni akarjuk (pl. időjárás jelentés). Ezt emberi

felolvasással hozzák létre. Az adatbázis hullámforma elemei (mondatok) tartalmazzák a beszédhangok legkülönbözőbb jellemző kombinációit és ezzel egyidejűleg a prozódiát is. Így – jó válogatás esetén – a prozódiát nem kell külön utólag ráültetni a hullámformára, az összefűzéssel egyidejűleg megjelenik az előállított beszéd hanghullámában. Az adatbázist precízen annotálni és címkézni kell hang, és szó szinten. A szintézis során a felolvasandó szövegnek megfelelő (általában szó, szókapcsolat, ill. mondatrész hosszúságú) hullámforma részeket válogatunk ki az adatbázisból, majd ezeket fűzzük össze, ideális esetben prozódiai módosítást végző jelfeldolgozás nélkül [30], [9]. Ez a terület képezi téziseim második csoportját.

1. táblázat. A kutatás során vizsgált gépi beszédkeltési módszerek áttekintése

Beszédszintézis módszer	**Prozódia előállítás**	**Beszéd adatbázis típusa**
„klasszikus" formáns szintézis (a kiindulási módszer)	szabály alapon, a kódoló vezérlő paramétereivel	parametrikus (formáns szűrő modell)
elemösszefűzéses (diád)	szabály alapon, hullámforma módosítással	hang, diád hullámforma elemek (logatomok)
elemösszefűzéses (triád)	szabály alapon, hullámforma módosítással	hang, triád és diád hullámforma elemek (logatomok)
elemkiválasztásos (korpusz)	indirekt, minta keresés alapú a mindenkori mondat időskáláján, jellemzően hullámforma módosítás nélkül	nagyméretű hullámforma adatbázis (felolvasásból) változó méretű elemekből (szó, szófüzér, mondat, stb.)
statisztikus parametrikus	statisztikus (HMM ill. DNN) modellel, amely paraméter n-gram alapján működik mondat szinten	parametrikus (LPC, harmonikus+zaj, szinuszos, stb.)
hullámforma-alapú statisztikus parametrikus (WaveNet/DNN)	statisztikus (DNN) modellel	neurális hálózat paraméterei (hullámformából tanítás és direkt generálás)

A gépi beszédkeltés terén az elmúlt években – számos előnyének köszönhetően – a statisztikai parametrikus beszédszintézis vált az egyik legaktívabb kutatási területté [10]. Ennek során először kinyerjük a jellemző paramétereket (például spektrális összetevők, alapfrekvencia, hangidőtartamok, hangok elhelyezkedése, hangkörnyezet) egy nagyméretű beszédkorpuszból, majd ezen paraméterek sokaságával modelleket alkotunk. Jellemzően a beszédfelismerésben már több évtizede sikeresen alkalmazott rejtett Markov-modell (HMM), valamint az újabban előtérbe

került Deep Neural Networks (DNN) alapú megközelítés a legelterjedtebb ebben a modellalkotásban. Ez a témakör fedi le téziseim harmadik csoportját.

A beszédtechnológia eredményeit egészen a 2000-es évek elejéig főleg csak unimodális módon (telefonos interakciók, felolvasás, beszédparancs értelmezés) alkalmazták. Ekkor kezdődött annak kutatása, hogyan lehet magas szinten tervezett ember-gép interakciókat mind grafikus, mind beszéd interfésszel megvalósítani [31], [32]. Ebbe a témakörbe esik téziseim negyedik csoportja.

Az 1. táblázatban foglalom össze a korábban felsorolt technológiákat két alapvető osztályozási szempont – a prozódia-előállítás és a beszéd kódolásának módja – szerint. A táblázatban szereplő (WaveNet/DNN) technológia a legújabb módszer, amelynek kutatásában elért kezdeti eredményekre dolgozatomban nem térek ki.

3. Kutatási célkitűzések

Az elmúlt 20 évben a számítástechnika technológiai fejlődése a gépi szövegfelolvasás területén is több, egyre összetettebb technológiai megközelítés kutatását és alkalmazását tette lehetővé, de kötelezővé is. Célom az, hogy megmutassam, hogy a gépi beszéd-előállítás témakörében az éppen aktuális technológia tükrében mindig változó kutatási kérdések merülnek fel. Ezek megoldása folyamatos kihívást jelent és egyrészt egymást követő alternatív tudományos generációkat eredményez. Másrészt az is jellemző, hogy a korábbi generációk nem avulnak el (mind a mai napig használatban vannak), hanem az újabb generációk más-más peremfeltételek között igénylik a működés optimalizálását és további alkalmazásokat tesznek lehetővé.

Kiinduló célkitűzésem a magyar nyelv sajátságait figyelembe vevő és kiaknázó gépi szövegfelolvasás olyan új modelljeinek és módszereinek kialakítása és ezekre épülő nemzetközileg is új interakciós lehetőségeknek a vizsgálata volt, melyek adott alkalmazási célokhoz jól illeszkednek. Munkám során több esetben magyar nyelvű beszédkorpuszokra támaszkodtam, azonban az értekezésben bemutatott megoldások jelentős részében nyelvspecifikus információt nem használtam fel.

Célkitűzéseim a következők voltak:

- *Célorientált gépi szövegfelolvasó rendszerek több generációjának kutatása, kialakítása és továbbfejlesztése elsősorban magyar nyelvre (I., II., III. téziscsoport).*

A legalapvetőbb technológiai korlátokat jellemzően az elérhető tárhely, az operatív memória és a számítási kapacitás jelenti. Ezek mellett kell a lehető legjobb gépi felolvasási minőséget elérni. A legjobb minőség azonban nem abszolút jellemző, hanem függhet a felhasználási körülményektől. Például a vak emberek számára az érthetőség a legfontosabb, de ezt közvetlenül követi a széles határok között (az átlagos 10-13 hang/s akár tízszereséig) állítható beszédsebesség, valamint a minél kisebb (akár 10 ms alatti) válaszidő. Ezzel szemben egy vasúti hangos utastájékoztató rendszerben jelentős háttérzaj mellett is érthető, kellemes hangzású bemondás a minőség meghatározója és akár több másodperces válaszidő is elfogadható. A kutatás célja ezeket a sokrétű felhasználói követelményeket a lehető legnagyobb mértékben kielégítő megoldások létrehozása. Ehhez számos új modellt, algoritmust és kutatási módszert kellett kidolgoznom.

- *Multimodális információs rendszerek hatékony megoldásainak kutatása (IV. téziscsoport, alkalmazások és műszaki alkotások).*

A gépi beszédkeltés gyakorlati felhasználásának elsődleges és kezdeti területe a távközlési alkalmazások voltak. Nem véletlen, hogy az egyetemek mellett ilyen cégek (Bell Laboratórium, NTT, stb.) finanszírozták az alapvető kutatásokat és hozták létre az első demonstrációkat. A másik

irány a személyi számítógépes, majd az okostelefonos alkalmazások területe, ahol sokáig a képernyő+billentyűzet volt a meghatározó interakciós eszköztár és csak speciális esetekben használták a beszéd modalitást (pl. képernyő felolvasó vak embereknek). Ezen a területen célom egyrészt az, hogy a sokszor csak angol nyelven elérhető rendszereket magyar nyelven is megvalósítsam, másrészt pedig arra törekszem, hogy az ismert megoldásokon túllépve, újszerű kombinációkat hozzak létre (pl. e-levél és SMS felolvasás). Ehhez szükséges új elvi megközelítéseket és tudományos eredményeket is kidolgoznom.

A beszéd nyelvfüggéséből természetszerűleg következik, hogy a különböző nyelvi változatok színvonalát nehéz összehasonlítani. Ezzel kapcsolatban elmondható, hogy a tesztjeink során kapott szubjektív minősítési értékek jellemzően a más nyelvekről megjelent publikációk értékei körül mozogtak. Ez azonban erősen függ az adott alkalmazási környezettől és az éppen összehasonlítás alatt levő rendszerektől.

Az I. téziscsoport szerinti eredmények alapján megalkotott ProfiVox diád/triád rendszert a magyar vak PC-s felhasználók jelentős része a mai napig jobban kedveli, mint a világcégek (Microsoft, Nuance, Google, stb.) mára elkészült magyar nyelvű változatait. A Jaws for Windows képernyőolvasó szoftver honosítási folyamatában pedig az amerikai, magyarul nem beszélő vezető fejlesztő e-levelében azt írta, hogy a magyar változatot az akkor mintegy 30 nyelvi változat közül a legjobb 3 között tartja számon.

A II. téziscsoport színvonalát jelzi, hogy a HMM TTS elmélet eredeti szerzői által jegyzett áttekintő cikk [10] az első megvalósítók között hivatkozik megoldásunkra. A III. téziscsoport eredményeit tartalmazó előadásunk [33] felkeltette a hasonló témán francia nyelven dolgozó kutatók figyelmét és érdeklődtek a részletek iránt.

A IV. (és a kapcsolódó II. és III.) téziscsoport eredményei kapcsán több H2020-as kutatási pályázat került benyújtásra és ezek közül kettő (PAELIFE és VUK AAL) támogatást nyert.

Kutatócsoportunk nemzetközi beágyazottságát az is jelzi, hogy a tématerület legjelentősebb konferencia sorozatán (Eurospeech, majd Interspeech) az 1989-es első alkalom óta a kétévente Európában tartott rendezvényen mindig volt legalább egy elfogadott előadásunk. 1999-ben mi rendezhettük meg az első nem nyugat-európai Eurospeech konferenciát. Azóta is az Európában legkeletebben tartott Eurospeech/Interspeech.

4. Eszközök és módszerek

Ebben a fejezetben a kutatásaim során használt adatbázisokat, eszközöket, azok működésének tesztelését, illetve a kutatási eredmények létrehozásának és kiértékelésének módszerét mutatom be.

4.1. A kutatás során használt adatbázisok

Beszéd-adatbázison a következőt értem: emberi beszéd hullámformája, az elhangzott beszéd fonetikai átirata és több szintű szegmentálási címkék párhuzamos halmaza. A beszéd-adatbázist (más néven beszédkorpuszt) jellemzően az adott kutatási feladathoz illesztve készítik el. Kutatásom során mindig az adott célnak megfelelő beszédkorpuszokat használtam, esetleg kombináltam.

A kutatás kezdetekor nem állt rendelkezésre célirányosan az elemösszefűzéses, az elemkiválasztásos (korpusz-alapú) és a statisztikus parametrikus szövegfelolvasó számára megfelelő magyar nyelvű beszéd-adatbázis, ezért először ezeket kellett kialakítani (2. táblázat). A táblázat adatbázisaiból a legfontosabbakat emelem ki.

Az elemösszefűzéses megoldás megalapozásához először a rendszertervet, majd a célhoz adaptált szövegadatbázist kellett megtervezni, majd annak felolvasásával a hangadatbázist is kialakítani (részletesebben az I. téziscsoport ismertetésében). Ezután következhetett a fejlesztői környezet [34] és a futtatható szintézis motor létrehozása. A diád, triád elemösszefűzéses beszédszintézishez annak elvi alapjait és korlátait figyelembe véve kellett megtervezni a felolvasandó szöveglistát (általában szó méretű értelmetlen hangsorok. Például: abáka, apáka, adáka....). Ezután került sor a felolvasásra, majd a diád, triád minták szegmentálására, címkézésére és kivágására. Így jött létre az első DIAD adatbázis, amit az igényelt hangkarakterek bővítésével követett a többi, majd a rendszer finomításával a TRIAD megoldás is (ld. 1. táblázat). A bővítés szükségességét a generált szintetikus beszéd minőségének folyamatos javítása hozta magával. A táblázatban feltüntettem a diádos teljes hanganyag időtartamát (kb. 28 perc „tiszta” időtartam, a stúdiófelvétel igénye több óra) és az ebből kézzel kivágott diád elemek (elemenként n*10 vagy n*100 ms) összegzett hosszát is (kb. 2,5 perc). Ebből a 2,5 perces hullámforma adatbázisból bármilyen tartalmú és hosszúságú beszéd előállítható (pl. egy könyv anyaga is felolvastatható) a megfelelő elemek összefűzésével. Érzékelhető a kézi feldolgozás munkaigénye is. Ilyen adatbázis 4-4 férfi és női hangra készült el.

Az elemkiválasztásos (korpusz) technológia kutatásához első lépésként az időjárás-felolvasás témakörét választottam, mivel korábbi kutatásaim során már felmértem annak komplexitását. Ez

volt az első ilyen magyar beszédadatbázis [9]. Ehhez először a megfelelő felolvasandó szöveglistát kellett létrehozni, majd azt felolvastatni és a felolvasott szöveget szegmentálni és címkézni. Így jött létre az IDO1 beszéd-adatbázis, ami több lépésben bővült a mostani méretre (ld. 2. táblázat). A kötetlen témakörré való kibővítés elősegítésére később fonetikailag kiegyenlített hanganyagot is felvettem ugyanezzel a beszélővel (FON1 beszéd-adatbázis). Az IDO1 adatbázissal kialakított korpusz-alapú modell működésének helyességét két másik témakörre elkészített rendszerrel ellenőriztem. A pályaudvari információszolgáltatás (PALYA1), valamint az árlista-felolvasás (ARU1) témakörében is hasonló szerkezetű adatbázisokat építettem ki. Szélesebb témakörű kísérleteimhez egy ügyfélszolgálati általános tematikájú teszt adatbázist (UGYF1) is létrehoztam.

2. táblázat. A kutatás során használt beszédkorpuszok

Tézis-csoport	Jel	Hangfelvétel hossza/adatbázis hossza (perc)	Nem	Nyelv	Célok
I.	DIAD1-DIAD4	28/2,5	férfi	magyar	Elemösszefűzéses (diád-triád) kutatások
	DIAD5–DIAD8	kb. 28 perc/2,5 perc	nő		
	TRIAD1	kb. 120perc/32 perc	férfi		
	TRIAD2	kb. 120perc/32 perc	nő		
II.	IDO1	630 perc	nő	magyar	Elemkiválasztásos (korpusz-alapú) kutatások
	FON1	100 perc			
	PALYA1	110 perc			Pályaudvari utastájékoztató kísérleti rendszer
	ARU1	330 perc			
	UGYF1	505 perc			
	SZAM1	10 perc/XXXX			Számok felolvasása 1 milliárdig
	SZAM2	10 perc/XXXXX	férfi		
	RADIO	516 perc	3 férfi		
III.	FON2-5	kb. 130 perc/fő	4 nő		Statisztikus parametrikus (HMM és DNN) kutatások
	FON6-10	kb. 130 perc/fő	5 férfi		
	BEA1	30 perc	nő		
	BEA2	31 perc	férfi		
IV.	GABOR	3 perc	férfi		Spemoticon kutatások

A prozódiai változatosság elemzéséhez rádiós hírekből, három férfi bemondó beszédéből is létrehoztam egy-egy beszéd-adatbázist (RADIO). Szintén felhasználtam erre a célra egy számfelolvasási célra korábban kialakított adatbázist (SZAM), [35].

A hanganyagok felmondását, rögzítését és a beszédkorpuszok kialakítását a BME-TMIT Beszédtechnológiai Laboratóriumának munkatársaival végeztük. A második téziscsoportomban ezeket az adatbázisokat használtam.

A harmadik téziscsoportban ismertetett statisztikus parametrikus témakörű kutatások lényegi célja, hogy sokféle beszédhangot és beszédstílust lehessen segítségével modellezni. Ezekhez a vizsgálatokhoz egyrészt felhasználtam az elemkiválasztásos kutatásokhoz kialakított adatbázisokat, másrészt ezeket újabb személyektől felvett fonetikailag kiegyenlített adatbázisokkal bővítettem. Ezeket kiegészítettem rövid (néhány szótagos) kijelentő és kérdő mondatokkal is. A spontán beszéd vizsgálatához felhasználtam a BEA adatbázis [36] két beszélőjétől származó felvételeket is.

A negyedik téziscsoport IV.1 altézis, valamint a III. téziscsoport megoldásaiban tetszőleges beszédszintézis, ill. beszédfelismerési technológia használható, ezért ezekhez nem kötődik adatbázis. A IV.2. altézisben az I. téziscsoport szerinti ProfiVox diádos/triádos technológia fejlesztői rendszerét alkalmaztam [8]. A kísérletekhez a GÁBOR hang diádos adatbázisát használtam fel.

4.2. A kutatások során felhasznált eszközök

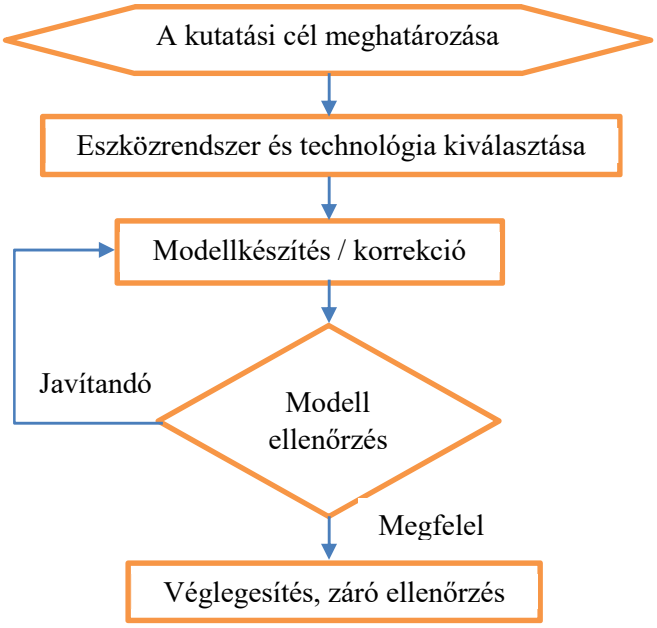
Kutatásaimhoz részben szabadon hozzáférhető eszközöket, részben pedig a BME-TMIT-en készült megoldásokat használtam. Ezek a következők:

- **VoXerver:** magyar nyelvű, automatikus beszédfelismerő beszéd-szöveg átalakítás, ill. kényszerített felismerés (Forced Alignment) üzemmódban. [37]
- **MVoxDev:** integrált szövegfelolvasó fejlesztői környezet [34]
- **Praat**: hullámforma elemzés és címkézés szoftver eszköze. [38]
- **HTS**: rejtett Markov-modell alapú gépi szövegfelolvasás keretrendszere [39]
- **DNN:** jellemző mély tanulási keretrendszerek, Merlin [40], Keras [41], Tensorflow [42], stb.

4.3. A kutatások módszertana

A kutatás jellemző módszerét az 5. ábra alapján mutatom be. A lépései a következők: az adott kutatási célhoz és az elérhető infrastruktúrához illeszkedő eszközrendszer és technológia kiválasztása, koncepció, modellkészítés, kis minta alapján koncepció ellenőrzés (pl. MOS/CMOS teszt néhány tesztelővel, ill. objektív mérések), annak alapján modell korrekció, a részletes végleges modell és rendszer kidolgozása, majd értékelése (MOS/CMOS teszt min. 20 tesztalannyal).

A módszereket befolyásolta a több évtizedes kutatás során bekövetkezett számítástechnikai technológiai fejlődés. A kutatás elején szinte kizárólag saját fejlesztésű szoftverekkel tudtunk dolgozni. A nyílt forráskódú és ingyenes keretrendszerek (pl. HTS és DNN eszközök) megjelenésével munkánk sokkal hatékonyabbá vált.



5. ábra. A kutatás módszere

A gépi szövegfelolvasás és a felhasználói felületek értékelésében általánosan elterjedt az eredmények MOS (Mean Opinion Score) és CMOS (Comparison Mean Opinion Score) alapú értékelése. Kutatásaim során én is ezen módszereket alkalmaztam. MOS alapú teszt esetén a tesztalanyok az elhangzott beszédet (mondat, szó stb.) 1-től (legrosszabb) 5-ig (legjobb) értékelhetik (egész számokkal),

CMOS esetén pedig jellemzően szintén 5 elemű skálán két minta közül kell a tesztalanyoknak eldönteniük, hogy melyik minta tesz jobban eleget a teszt osztályozási kritériumának (például minőség, természetesség, érthetőség). A tesztek során bizonyos esetekben a „minőség” fogalom

értelmezését a tesztalanyokra bíztam. Ekkor az osztályzás általános visszajelzést ad arról, hogy a tesztalanyok mennyire tartják jónak vagy rossznak az adott rendszert. Ez esetben a rendszer értékelésében számos paraméter, például természetesség, érthetőség, a hang által a tesztalanyban keltett érzelem, stb. szerepet játszik. Egyes esetekben arra kértem a tesztalanyokat, hogy például a bemondás természetességét osztályozzák. A MOS és CMOS típusú meghallgatásos teszteken elért pontszámok átlagát grafikonon, illetve oszlopdiagramon ábrázoltam.

Az utóbbi években áttértünk a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor [43]) teszt módszeralkalmazására is, mert kevesebb tesztelő személlyel lehet statisztikailag értékelhető eredményekhez jutni. Itt egy 0-100 közötti skálán kell értékelni a mintákat. Az értékelést segíti, hogy a rendszernek része egy (rejtett) 100%-osnak számító referencia és ahhoz képest kell a tesztmintákat értékelni.

5. A diád és triád elemek összefűzésén alapuló gépi szövegfelolvasás (I. téziscsoport)

A formáns-alapú rendszerek érthető, de robotos hangminőséget állítottak elő. Ez rövid üzenetek meghallgatását lehetővé tette, de hosszabb szövegek felolvasása jelentős kognitív terheléssel járt. A látássérült emberek számára különösen nagy nehézséget okozott a felolvasó rendszerek egész napos használata. A szakirodalomban felmerült, hogy emberi beszéd rögzítésén alapuló megoldással előrelépést lehet elérni [6]. A minőség javításán túlmenően fontos szempont volt a látássérültek számára, hogy a beszéd érthető maradjon széles tartományban felgyorsított beszédsebesség mellett is.

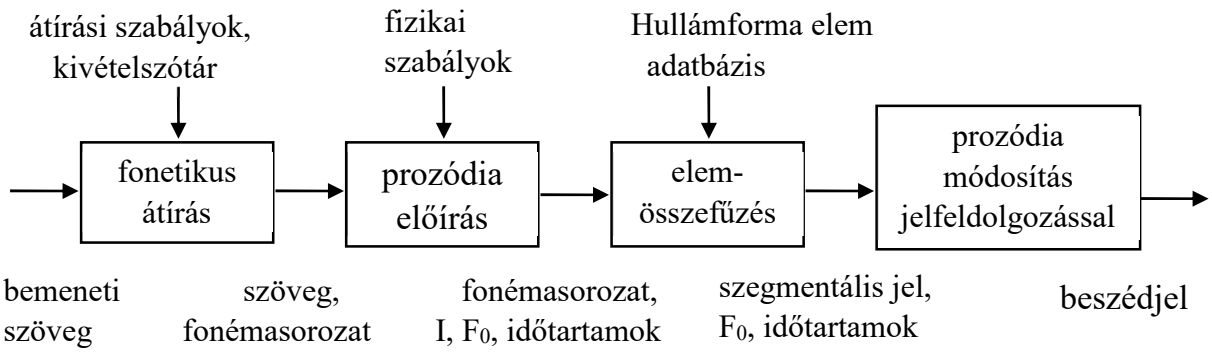
A fenti szempontok figyelembe vételével kidolgoztam az első magyar nyelvű hullámforma elemösszefűzéses gépi szövegfelolvasó rendszertervét. Megterveztem a diád és triád hullámforma elemek megvalósításához felhasználható akusztikus adatbázis szerkezetét és az annak elkészítéséhez szükséges szövegkorpuszt. Munkatársaimmal megvalósítottuk a rendszert és több hangra, valamint német nyelvre is kiterjesztettük. Célorientált megközelítéssel optimalizáltam, és adaptáltuk látássérültek kommunikációját segítő képernyőolvasó rendszerhez, amely ma a legelterjedtebb PC-alapú megoldás Magyarországon (a Jaws for Windows-t több ezer látássérült ember használja, a Robobraille szövegből hang fájlkonverziós szolgáltatás pedig bárki számára ingyenesen igénybe vehető).

5.1. A diád és triád elemösszefűzéses gépi szövegfelolvasó eljárás *(* I.1 tézis)

Kidolgoztam a magyar nyelv sajátosságainak megfelelő első diád és triád hullámforma elemösszefűzéses gépi szövegfelolvasó eljárás rendszertervét (ld. 6. ábra), amely diád és triád méretű magyar hangkapcsolódások felhasználásával készít gépi beszédet, és igazoltam, hogy az ezek felhasználásával létrehozott rendszer MOS (Mean Opinion Score) szubjektív értékelés szerint jobb hangminőséget ad, mint a korábbi, más elven működő megoldások (például Hungarovox [28]*, Brailab* [44]*, PC talker* [45]*) Az eljárást kiterjesztettem német nyelvre is.*
Alátámasztó irodalmak: [8], [1]

A korábbi parametrikus (elsősorban formáns-alapú, szabályokon alapuló), de erősen robotos hangzású gépi szövegfelolvasási technológia továbbfejlesztésére a 80-as évek végétől alakult ki az a koncepció [6], [46], [47], hogy próbálkozzunk természetes beszéd rögzítésével, címkézésével és a visszajátszáskor a megfelelően kiválasztott hullámforma elemek összefűzésével és (ha

szükséges) jelfeldolgozás segítségével történő dallam ráültetéssel. Ennek egyik lehetséges megoldását a 6. ábra mutatja be.



6. ábra. Hullámforma elemösszefűzésen alapuló beszédszintetizátor egy lehetséges modellje. [8]

A formáns alapú parametrikus szintézishez képest az előrelépés alapja az, hogy emberi beszédből hozzuk létre a hullámforma adatbázist, így az természetszerűleg magában hordozza az emberi jelleget. A modell előnye, hogy a fonetikus átírást és a prozódiát tervező modulok lényegileg átvehetők a korábbi kutatási eredményekből. Hátránya viszont, hogy a hullámforma elemek megfelelő címkékkel való ellátása (annotálása), és prozódiai jellemzőiknek (alapfrekvencia, időtartam és intenzitás) módosítása a formáns modell megoldásához képest lényegesen bonyolultabban valósítható meg. Erre az I.2 tézis tárgyalása során térek ki.

A magyar beszéd 39 beszédhanggal (24 mássalhangzó (C) és 14 magánhangzó (V) plusz a szünet (_ jel) lefedhető (ld. 3. táblázat) úgy, hogy a hosszú mássalhangzókat a rövid változatukból jelfeldolgozással állítjuk elő. A magánhangzók minőségére nagyon érzékenyek vagyunk, ezért célszerű a rövid és a hosszú változatokat külön kezelni. A hosszú mássalhangzókat csupán időtartam módosítással tudjuk előállítani. Különlegesség a zöngétlen palatális réshang (zöngétlen j, 39-es számkód), amely külön elemként került meghatározásra (lépj) ugyanis ezt a hangot írásunk nem jelöli. A dz és dzs ellenben megfelelő adatbázis tervezéssel jó minőségben megoldható a többi elem felhasználásával, ezért nem tekintjük külön hangnak a szintézis szempontjából. A hullámforma összefűzés módszerének egy lehetséges megoldása az, hogy alapelemnek hangpárokat reprezentáló beszéd hullámforma részleteket (ún. diádok, angolul diphone) választunk. Ekkor például az „alma” szót *_a, al, lm, ma, a_* (_ a szünet jele) diádokból lehet előállítani. Itt a magánhangzók ketté vannak vágva, a hangnak csak a fele szerepel a diádban. Hanghármasok (triádok) esetében az „alma” szót az *_al* és *ma_* triádokból valamint az *lm* diádból tudjuk előállítani. A magyar diád adatbázisban tehát mintegy 40^2=1600, a németben pedig (ld. 4. táblázat) 50^2=2500 elemre van szükség.

3. táblázat. A magyar nyelvű ProfiVox diád-triád alapú beszédszintetizátor által kezelt beszédhangok készlete

Szám kód	Magyar karakter	Példa	Saját karakter jel	IPA Unikód szám	IPA Ascii szimbólum
1	(szünet)		_		
2	á	láb	A:	0250 02D0	a&:
3	a	hat	a	0254	c&
4	o	sok	o	006F	O
5	u	fut	u	0075	U
6	ü	süt	U	0079	Y
7	i	hit	i	0069	I
8	é	méz	E:	0065 02D0	e:
9	ö	köt	O	00F8	o/)
10	e	vet	e	025B	E
11	b	bál	b	0062	B
12	p	tár	p	0070	P
13	d	dán	d	0064	D
14	t	tár	t	0074	T
15	g	gát	g	0067	G
16	k	kád	k	006B	K
17	gy	gyár	G	025F	j-
18	ty	tyúk	T	0063	C
19	m	már	m	006D	M
20	n	nád	n	006E	N
21	ny	nyom	N	0272	nj)
22	j	jön, lyuk	j	006A	J
23	h	hát	h	0068	H
24	v	vád	v	0076	V
25	f	fát	f	0066	F
26	z	zár	z	007A	Z
27	sz	szép	s	0073	S
28	c	cél	c	0074 0073	Ts
29	zs	zsír	Z	0292	3"
30	s	só	S	0283	S
31	cs	cső	C	0074 0283	TS
32	l	láp	l	006C	L
33	r	rák	r	0072	R
34	ó	pók	o:	006F 02D0	o:
35	ú	kút	u:	0075 02D0	u:
36	ű	fűt	U:	0079 02D0	y:
37	í	szít	i:	0069 02D0	i:
38	ő	sőt	O:	00F8 02D0	o/):
39	j*	kapj	j	006A	J

4. táblázat. A német nyelvű ProfiVox rendszer által kezelt beszédhangok készlete

Számkód	Saját karakter jel	Példa	Saját karakter jel	Saját karakter jel	Példa
1	_	(pause)	26	z	singen
2	a:	sah	27	s	Kasse, daß
3	oi	Bäume, deuten	28	ts	Zahn, einzeln
4	o:	holen	29	Z	Garage
5	u:	gut, Uhr	30	S	schon
6	ü:	führen	31	tS	Rutsch
7	i:	hier	32	l	Lampe
8	e:	Tee, geben	33	r	rennen, Herr
9	ö:	hören	34	6	räder, Vater
10	ö	Böll	35	au	Auto, laut
11	b	bunt	36	ai	mein, Ei
12	p	Punkt, Klappe	37	x	auch, doch
13	d	danke	38	c	ich, Milch
14	t	Tier, bitte	39	o	Ochs, voll
15	g	gut, gegen	40	@	Liebe, meine
16	k	kein, Kuckuck	41	an	Chanson
17	E	essen	42	a	falsch
18	E:	ähnlich	43	e	degeneriert
19	m	Mutter	44	on	Chanson
20	n	nein	45	u	kulant
21	ng	singen	46	ü	füllen, fünf
22	j	ja	47	I	Fisch
23	h	Haus	48	pf	Tropfen
24	v	Wagen	49	|	(glottal stop)
25	f	fangen, Vater			

Ez az adatmennyiség mind kézi feldolgozási igény, mind a 90-es évek végén elérhető PC-s tárhely és számítási kapacitás szempontjából reálisnak számított. A technológia fejlődésével kiderült, hogy a kis erőforrás-igényű diád alapú beszédszintetizáló rendszerek alkalmasak voltak a 2000-es évek elején megjelenő okostelefonokon valós idejű működésre, amivel az akkori beszédtechnológiai csúcstechnológiát hoztuk létre, és a vezető ipari partnerek (pl. MATÁV, Westel és Nokia) érdeklődését is felkeltettük (ld. 9.1, 9.2 és 9.3 fejezet). Ma is napi felhasználásban van, különösen vak emberek számára fejlesztett PC-s és mobiltelefonos alkalmazásokban.

Számszerű kiértékelés

A rendszert a MAILMONDÓ szolgáltatás [48] és [49] fejlesztése és alkalmazása során széles körben teszteltük és megállapítottuk, hogy jobb minőséget nyújt, mint a korábbi magyar nyelvű gépi szövegfelolvasó megoldások (ld. 11. ábra, 32.o). A német nyelvű változatot kutatási együttműködés keretében a TU Kaiserslautern és a Fraunhofer IESE anyanyelvű munkatársaival validáltuk [50].

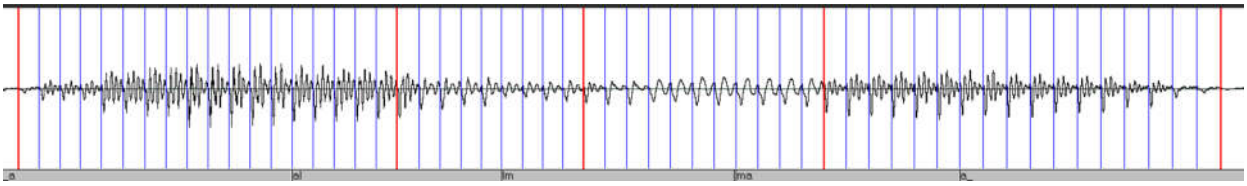
Konklúzió

A tézisben ismertetett kutatás eredményeként a korábbi, a gyakorlatban szinte kizárólag fogyatékos emberek számára alkalmazható megoldásokon túlmutató, közcélú távközlési szolgáltatásokba is bevezetett beszédszintetizáló rendszer jött létre.

5.2. Diád és triád alapú rendszerek beszédadatbázisa (I.2. tézis)

Megterveztem az első magyar diád és triád hullámforma elemek megvalósításához felhasználható magyar nyelvű felolvasásos beszédadatbázis szerkezetét és az annak elkészítéséhez szükséges, az az átlagos prozódiai jellemzőket biztosító szövegkorpuszt. Alátámasztó irodalmak: [51], [1]

Az adatbázis tervezése során figyelembe kellett venni, hogy a tervezés időpontjában érvényes számítástechnikai korlátok mellett (memória és CPU) egy hangkapcsolati egységhez csak egyetlen adatbáziselem tartozhatott. A prozódiai megvalósítással kapcsolatosan fontos felismerés volt, hogy a hangkapcsolati egységeknek átlagos prozódiai jellemzőkkel (átlagos alapfrekvencia, időtartam és intenzitás) kell bírniuk, hogy a prozódiai módosítást pozitív és negatív irányban is viszonylag kis torzítás mellett el lehessen végezni rajtuk. A prozódiai módosítás megvalósításához a zöngés adatbáziselemeket periódusonként pontos címkézéssel (periódus határ) kellett ellátni. Úgy határoztam meg a jelölést, hogy a periódus kezdete egy (lehetőleg) kis energiájú, pozitívból negatívba váltó nulla átmenetnél legyen, a vége pedig az ellenkező irányú nulla átmenet váltásnál. Ezzel a megoldással elértem, hogy a periódushatár egyben a diád vagy triád elem határa is lehet.



7. ábra. Az „alma” szó 5 diádja látható.
A hangperiódusok határa kék színnel, a hanghatár piros színnel van jelölve, a diádok határát az alsó szürke sávban levő jelölések melletti vonalak jelölik.

A megoldás eredményét a 7. ábra illusztrálja. A zöngétlen szakaszokon az adott beszélő jellemző átlagos zöngeperiódus idejének megfelelő fix értéket (férfi hangnál mintegy 10 ms, női hangnál mintegy 5 ms) „virtuális” periódushosszt alkalmazunk. Zárhangoknál célszerűen a zár kezdete és vége a hanghatár.

A felolvasandó szövegkorpuszt úgy kellett kialakítani, hogy a felolvasása után létrehozott hullámforma állományból optimális minőségben és közel egyenletes alapfrekvenciával lehessen kivágni az adatbáziselemeket. A magyar beszéd szintetizálásához a fentebb említettek szerint 14 magánhangzót és 24 mássalhangzót felhasználva az 5. táblázat szerinti diádokra van szükség. A triádok esetében elsősorban a CVC kapcsolatok megvalósítása célszerű, a magánhangzók közepén történő vágás okozta torzítás kiküszöbölése miatt. Ekkor azonban a szükséges elemszám 10.000 fölé nő (a magyar változatban 25x14x25=8750 triád + 1520 diád, a német változat pedig ennél is nagyobb), ami mind az adatbázis tervezését, mind megvalósítását illetően jelentős többletterhet jelent a diádos megoldáshoz képest.

5. táblázat. A magyar nyelv szintéziséhez szükséges diád változatok darabszáma
(_ a szünetet jelöli)

Hangkapcsolat típusa	CV	VC	CC	VV	_V és V_	_C és C_	Összesen
Darabszám	336	336	576	196	28	48	1520

A felolvasandó szöveget célszerű úgy kialakítani, hogy a CV és VC szerkezetű diádok magánhangzóinak spektrális szerkezetét minél kevésbé befolyásolja a szomszédos hangok hatása (koartikuláció). Korábbi fonetikai vizsgálatokból ismert [52], hogy a *k* hang kevéssé befolyásolja a megelőző és a követő hangok frekvenciaszerkezetét. Emiatt választottuk ezt a hangot a diád hangjait megelőző, ill. követő hangnak. Az ezeket az elemeket közrefogó magánhangzónak pedig az *a* hangot választottuk, mivel artikulációja egyszerű. Az így kialakított mesterséges szavak (logatomok) együttesét nevezzük elemiszöveg-halmaznak.

6. táblázat. A szövegelemek felépítési elve

Megvalósítandó diád típus	VC	CV	VV	CC
A szövegelem felépítése	a+k+VC+a	a+CV+k+a	a+k+VV+k+a	CC hosszabb hangsorban
Mintapélda	ak*ab*a	a*ba*ka	ak*aá*ka	ha*mv*asodik

A CC kapcsolatokban a vágás helyén előforduló esetleges illeszkedési hiba kevésbé zavaró, mint a magánhangzóknál. Viszont a természetes ejtéshez közeli szerkezet fontos, ezért ezekhez a diádokhoz hosszabb, a természetes nyelvben is előforduló szövegelemeket választottunk. A szövegelemek szerkezetét és példáit láthatjuk a 7. táblázaton.

A fenti elvek szerint kialakított szöveges adatbázist strukturált, jól olvasható állományba rendeztük, ami segítette a szöveget felolvasó személy (bemondó – voice talent) munkáját a stúdiófelvétel elkészítése során. Az egyenletes minőségű bemondáshoz egyedi módszertant alakítottunk ki (rögzített szájtávolság a mikrofontól, minimális mozgás a felolvasás közben, egyenletes hangmagasság tartása, egyforma szünetek a logatomok között stb.). A bemondó jellemzően egy, legfeljebb két oldalnyi szöveget olvasott fel egyszerre (ez került egy hangfájlba). A számos hangfájlból félautomatikus ellenőrzési és szerkesztési módszerek segítségével állnak elő a köztes állományok. A végleges adatbázis titkosítási, verziókövetési és memória optimalizálási megoldások alkalmazásával jön létre. A diádos adatbázis mérete 22 kHz 16 bit mintavételezés esetén beszélőnként átlagosan 6,3Mbyte, a triádos adatbázis pedig jellemzően 90Mbyte körül van. Az adatbázis elkészítéséhez és több iterációs kör után történő végleges kialakításához az MVoxDev fejlesztői rendszert használtam [53].

Számszerű kiértékelés

A rendszert a II.1 tézisben ismertetett teszteléseknek vetettem alá és megállapítottam, hogy a formás szintézis alapú megoldásnál jobb minőségű (az 1-5-ös skálán 1,5-el) gépi beszédet szolgáltat (ld. 11. ábra, 32.o.). A folyamatos továbbfejlesztések (közel 20 év) alatt számos férfi és női hangkarakter került kialakításra. Ezek közül a Jaws for Windows rendszer képernyőfelolvasó magyar hangjaként alkalmazott ProfiVox változatban négy hang (két férfi és két női) érhető el. A Robobraille szolgáltatás (https://www.robobraille.org/hu/szoveg-konvertalasa) pedig egy-egy férfi ill. női hangot támogat.

Konklúzió

A rendszer a központi szerveren futtatható változaton túlmenően (MailMondó szolgáltatás), a világon először mobiltelefonon futó szolgáltatás részeként (SMSmondó ill. SMSRapper az angol változat, ld. 9.1 fejezet) [54], [55] is elérhetővé vált. Az adatbázisok optimalizálásával (mintavételi frekvencia, beszédminta kódolás, gyakoriság figyelembe vétele, stb.) és gyors prozódia módosító algoritmus kidolgozásával elértük, hogy a beszédsebesség széles határok között változtatható, ami kritikus funkció a látássérült emberek számára. Napjainkban (2019) jutottunk el oda, hogy ez a technológia PC-s képernyőolvasóba integráltan minden magyar látássérült ember számára ingyenesen hozzáférhető lett, egyelőre 1 év időtartamra [56] . Folyamatban van a legnagyobb magyar banknál is a rendszer több száz pénzkiadó automatába (ATM) történő üzembe helyezése, így a pénzfelvétel gépi beszéddel történő megkönnyítése látássérült emberek számára is lehetővé válik.

6. Célorientált, korpusz-alapú gépi felolvasó rendszerek (II. téziscsoport)

A 90-es évek második felében kezdett megfogalmazódni az a koncepció, amit korpusz-alapú beszédszintézisnek nevezünk [30]. Az elképzelés alapötlete abból az általánosan elfogadott elvből fakad, hogy egy hullámforma-összefűzésen alapuló szövegfelolvasó rendszer minőségét döntően a *beszédadatbázisban szereplő egyazon időben ejtett elemek hossza határozza meg*. Minél hosszabb egybetartozó hullámforma elemekből állítjuk elő a szintetizált beszédet annál jobb lesz az elért minőség. Tehát az elemösszefűzéses megoldással szemben, ahol egyrészt egy-egy hangkapcsolat (diád és/vagy triád) egy vagy több realizációja az alapelem, az elemkiválasztásos esetben hosszabb elemekből építkezünk. Az ideális tehát az lenne, ha minden lehetséges felolvasandó szöveg, de legalábbis minden lehetséges mondat szerepelne elemként a rendszer adatbázisában. Természetesen ez a gyakorlatban kivitelezhetetlen, ezért olyan egységeket rögzítenek az adatbázisba, hogy a szintetizálandó mondat nagy valószínűséggel hosszú elemekből legyen összefűzhető.

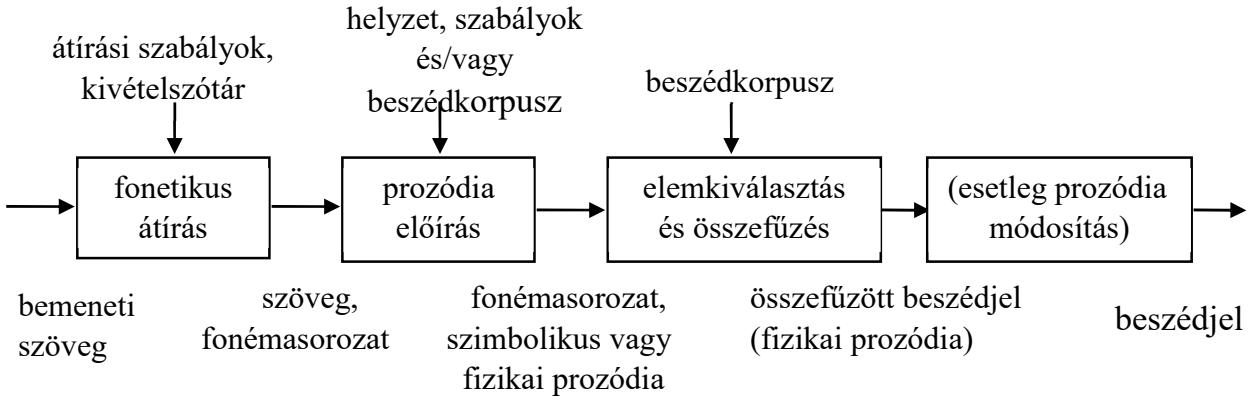
Három célorientált alkalmazási területhez (időjárás-jelentés, árlista és pályaudvari tájékoztató felolvasása) adaptáltam a rendszert. Megmutattam, hogy ennek a technológiának a felhasználásával lehetséges az emberi felolvasáshoz megtévesztésig hasonló magyar nyelvű gépi felolvasást létrehozni.

6.1. Magyar nyelvű korpusz-alapú gépi szövegfelolvasás modellje (II.1. tézis)

Kidolgoztam magyar nyelvre az első korpusz-alapú hangnyomás-idő függvények automatikus válogatásán alapuló gépi szövegfelolvasó eljárás modelljét, amely szavak, szókapcsolatok, mondatrészek hangnyomás-idő függvényeinek célorientált összefűzésével készít gépi beszédet, valamint az ehhez kapcsolódó, fonetikai szempontok szerint kialakított költségfüggyényeket és indirekt prozódiai modellt. MOS vizsgálatokkal igazoltam, hogy jobb hangminőséget eredményez, mint az I. téziscsoport szerinti megoldások. Alátámasztó irodalmak: [57], [58], [1]

Létrehoztam az első magyar nyelvű korpusz-alapú, hullámforma elemválogatásra épülő gépi szövegfelolvasó rendszer modelljét (8. ábra). Jellemzően témaspecifikus adatbázis készül, viszont a megoldás alkalmas tetszőleges szöveg felolvasására is, viszont ebben az esetben a hangminőség változó lehet. A prozódiát jellemzően nem utólagos módosítással állítjuk elő, hanem a 9. ábra (26.o) szerinti modellt alkalmazva. Ez azt jelenti, hogy a számos adatbáziselemből olyan elemet választunk ki, ami az adott mondat adott hangsorának megvalósításához szükséges prozódiai jellemzőkkel bír. Ha ilyen elem nincs, akkor (kivételesen) kerül sor a leginkább illeszkedő elem kiválasztására és annak jelfeldolgozással történő prozódiai módosítására (pl. kijelentő mondat végén az alapfrekvencia csökkentésére). A megoldás elvét a 8. ábra mutatja.

Ennek az elvnek egy régi, legegyszerűbb megoldása az ún. kötött szótáras beszédszintetizáló rendszer, mint például az autóbuszokon alkalmazott bemondások digitális rögzítése, majd megfelelő egyszerű vezérlés (pl. nyomógombok) segítségével történő visszajátszása. Például: A következő megálló ---- Keleti pályaudvar. A mondat első fele a rögzített elemet képviseli, a mondat második eleme a változót. Fontos látni, hogy az ilyen összeillesztéseknél a prozódiának illeszkednie kell egymáshoz. Ez a példában azt jelenti, hogy a rögzített rész mindig az üzenet kezdete, a megálló neve pedig a vége (ha megcserélnénk a kettőt, és úgy játszanánk le, akkor prozódiailag természetellenes hangzást kapnánk, amire mindenki felkapná a fejét.). Természetesen ennek a kötött szótáras megoldásnak egyrészt jelentős a tárigénye, másrészt erősen korlátozott a témaköre.

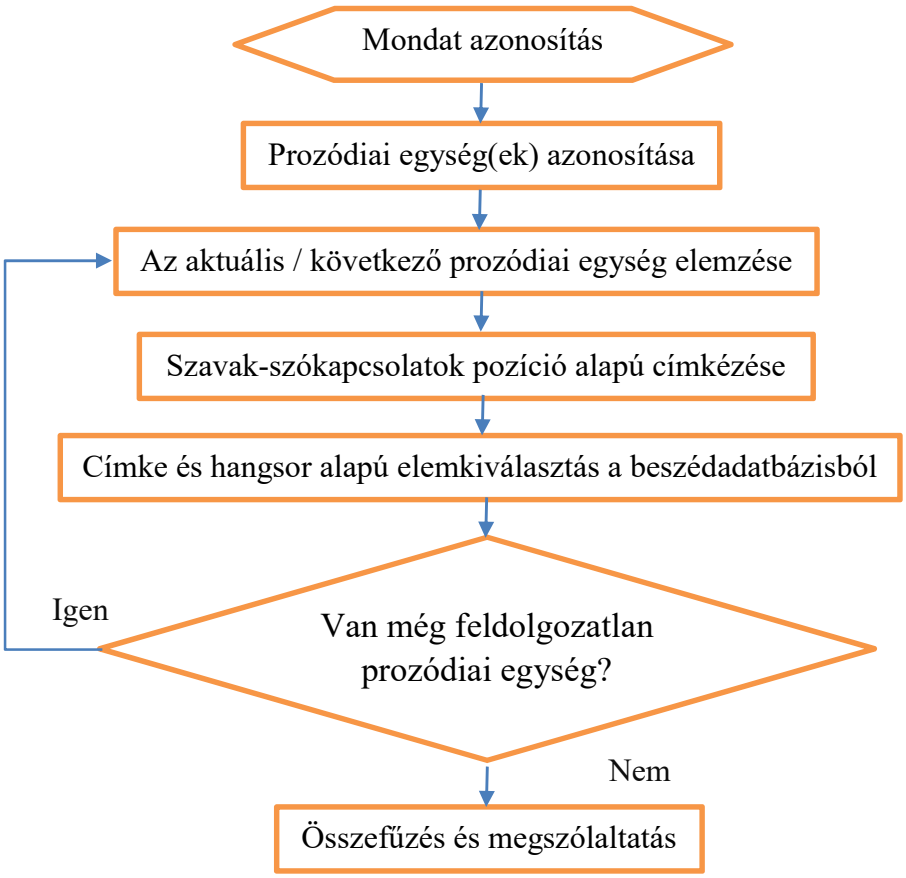


8. ábra. Korpusz alapú, hullámforma elemkiválasztásos beszédszintetizátor modellje. [59]

A fenti koncepció alapján külföldön már készült néhány korpusz-alapú beszédszintetizátor a világnyelvekre [30], [60], magyar megoldást azonban kutatócsoportommal elsőként hoztam létre. Munkánk során felhasználtuk a korábbi magyar nyelvű kutatások [35] eredményeit is. Kutatásaink során arra a fő kérdésre kerestük a választ, hogy lehetséges-e olyan gépi beszédkeltési modellt létrehozni magyar nyelvre, ami akár az emberi bemondásra megtévesztésig hasonló kimenetet tud létrehozni kötött, de nagy változatosságot tartalmazó tématerületen. A más nyelvekre kidolgozott modellek nem feltétlenül hasznosíthatók, hiszen a magyar nyelv ragozó jellege miatt például az angol nyelvre kidolgozott szó-alapú megközelítések nem alkalmazhatók közvetlenül.

Az elemkiválogatás alapú beszédszintetizátorok két legfontosabb eleme a beszédkorpusz, valamint az abból automatikusan válogatást végző algoritmus.

A kutatásunk eredménye szerint a beszédkorpuszt két szempont szerint kell összeállítani. Az egyik, hogy a témakörnek megfelelő szavakat, szófüzéreket tartalmazza. A második pedig, hogy a szavak, szófüzérek, mondatrészek, mondatok korrekt prozódiájának előállításához újfajta indirekt modellt kell megalkotni. Ez utóbbi végleges rendszertervét elkészítettem és alább ismertetem. A szintézis során az alapegységnek a mondatot tekintem, tehát mondatnyi egységeket szintetizálunk egy menetben. A modell szorosan összefügg a szintetizálandó szöveg szerkezetével, döntően kijelentő mondatok előállítását támogatja. A kijelentő mondat prozódiai szerkezete jól körülhatárolható, ismert egységekből áll [61]. A mondat szavait a mondaton belüli hely szerinti pozicionálással (hol van a szó a mondatban), valamint a központozással (vesszők, gondolatjelek stb.) hoztam kapcsolatba. A modell lényegét a 9. ábra mutatja be.



9. ábra. Az indirekt prozódiai modell feldolgozási folyamata

A modell működését az alábbi mintamondaton mutatom be:

Szombaton egynapos enyhülés következik, változó napsütéssel, helyenként záporokkal.

Az első prozódiai egység: *Szombaton egynapos enyhülés következik,* (PR1)
Az elemzés eredménye:
PR1(k)/KSZ *Szombaton PR1(k)/BSZ egynapos PR1(k)/BSZ enyhülés PR1(k)/ZSZ következik*
A második prozódiai egység: *változó napsütéssel,* (PR2)

Az elemzés eredménye:
PR2(b)/KSZ *változó PR2(b)/ZSZ napsütéssel*
A harmadik prozódiai egység: *helyenként záporokkal* (PR3)
Az elemzés eredménye:
PR3(z)/KSZ *helyenként PR3(z)/ZSZ záporokkal.*

A címkék jelentése:
PRx: az x. prozódiai egység
(k): mondatkezdő prozódiai egység
(b): mondat belsejében elhelyezkedő prozódiai egység
(z): mondatzáró prozódiai egység
KSZ: kezdőszó a prozódiai egységen belül
BSZ: belső szó a prozódiai egységen belül
ZSZ: zárószó a prozódiai egységen belül

A modellel elvégeztem a prozódiai címkézést a szöveges adatbázisban, a beszédadatbázisban és a szintetizálandó mondatban is. A szintéziskor a szó, szófüzér kiválogatása során a prozódiai címkék szerinti egyezést keressük. Erre külön válogató függvény szolgál. A modell alkalmazásával az esetek nagy részében nincs szükség prozódiai jellegű jelfeldolgozás használatára a szintézis során, mert anélkül is megfelelő minőség érhető el.

A szintézis optimális akusztikai alapegységének a szó elemet választottam. Ennek megfelelően alakítottam ki a beszédkorpusz elkészítéséhez szükséges felolvasandó szövegtárat. A szó méretű elem egyrészről hosszabb a diád-triád elemeknél, tehát akusztikai tartalma biztosan jobban képviseli az optimális hullámformát, másrészről az ember percepciós rendszere inkább a szó feldolgozására épül, mint a hangokéra, vagy a hangkapcsolatokéra. Ha tehát jó akusztikai tartalmú és prozódiájú szó kerül a szintetizálandó mondatba, akkor természetesebb hangzásúnak fogjuk ítélni, mint a diád/triádokból összerakott ugyanazon szintetizált egységet. Mindezekből adódik, hogy a szintézishez használt beszédadatbázisnak minden szóból legalább háromfélét kell tartalmaznia (mondat kezdő, -belső helyzetű és -záró elem).

Az elemkiválasztás és összefűzés modulban két költségfüggvény összegének minimalizálása valósul meg *új, fonetikai szempontok szerint kialakított költségfüggvények* alapján. Az egyik költségfüggvény (*C(n)*) az egyes (szó és hang szinten potenciálisan eltérő) elemek összefűzésének költsége (ún. concatenation/összefűzési költség) amit az elemek egymáshoz illeszkedése/folytonossága alapján származtatok. A másik költségfüggvényt (*P(n)*) annak alapján származtatom, hogy hangsor és hangkörnyezet szempontjából a kiválasztott elem (szó, szófüzér vagy hang) mennyire felel meg a prozódiai követelményeknek (Prosodic target/prozódiai illeszkedés).

A *C(n)*) függvényt az alábbiak szerint definiáljuk:

$$C(n) = \sum_{i=1}^{K} w_n(i) * D(u_n(i), u_n(i+1)) \qquad (1),$$

ahol

C(n) a K+1 elem összefűzéséből előálló n-dik alternatíva összefűzési költsége,

$u_n(i)$ az i-dik összefűzött elem az n-dik alternatívában,

$D(u_n(i), u_n(i+1))$ két egymás követő elem összefűzési költsége,

$w_n(i)$ az n-dik alternatívában az i-dik és az i+1-dik elem összefűzési költségének súlytényezője.

Mivel a kiejtés folyamatos, a (szó)határokon törekedni kell arra, hogy a spektrális illeszkedés (pl. formánsmenet) is folyamatos legyen. A szavak első és utolsó hangjának illeszkedését vizsgálom, és az illeszkedés költségét több szempont alapján számítom ki. Magas költségű például, ha a szóhatáron magánhangzók találkoznak (dunántúli áramlások). Az ilyen szavak magas költséget képviselnek. Nulla a költség, ha a két szó egymás mellett helyezkedik el a beszéd-korpuszban, hiszen ekkor a csatlakozásuk is optimális. Ebből adódik, hogy akkor nagyon optimális a keresés, ha nem szavakat, hanem szófüzéreket találunk a korpuszban. Az esetek nagy részében (ha a beszéd-korpusz elég nagy) ez meg is valósul, így a szintetizált szöveg hangzása közel lesz a természeteshez. Felhasználtam többek között azt a kutatási eredményt is, hogy azonos képzési helyű mássalhangzók akusztikai megvalósulása hasonló átmeneti fázisokat okoz a hozzájuk csatlakozó magánhangzóban [62], továbbá a mássalhangzók képzési módjának osztályozását és a gerjesztés fajtáját (zöngésség-zöngétlenség). A mássalhangzók képzési helyéből adódó azonos akusztikai vetületeket a 10. ábra mutatja be.

Az optimális összefűzési pontokat elsősorban a 10. ábra szerinti 7 artikulációs vetületi sor, illetve a beszédjel energiája dönti el. Nem célszerű összeillesztést végezni nagy energiájú jelszakaszban (például magánhangzóban), a kis energiájú helyeket kell előnyben részesíteni. Szabad illeszteni a hangsor minden olyan pontján, ahol gerjesztésváltás megy végbe (tiszta zöngés szakaszt tiszta zöngétlen követ és fordítva, itt ugyanis a jelben intenzitás minimum keletkezik), továbbá a hangok belsejében lévő néma fázisokban, illetve zöngeszakaszokban Ha tehát az akusztikai vetület ugyanaz, és például gerjesztésváltás van a két elem határán, akkor az összeillesztési költség értéke kicsi lesz, hiszen a spektrális folytonosság biztosított és az illesztésnél kicsi az energia.

	zárhangok							zár-rés hangok				réshangok						nazálisok				
	b	p	d	t	gy	ty	g	k	c	dz	cs	dzs	v	f	z	sz	zs	s	h	m	n	ny	j	l	r
két ajak	☒	☒																		☒					
ajak-fog													☒	☒											
fog-fogm.			☒	☒					☒	☒					☒	☒					☒				
fogmeder											☒	☒					☒	☒						☒	☒
e.szájpadl.					☒	☒																☒	☒		
h.szájpadl.							☒	☒																	
gége																			☒						

10. ábra. A magyar mássalhangzók képzési hely és mód szerinti csoportosítása. Az ugyanazon sorban lévő mássalhangzók hasonló akusztikai vetületet hoznak létre a hozzájuk csatlakozó magánhangzóban [62]

Hasonló elvek alapján kialakult az a fonetikai szabályrendszer, amellyel ki lehet jelölni a vágás konkrét helyét (a vágási pontot). Erre mutat példát a 7. táblázat. Itt szempont az is, hogy a kiválasztott elem a mondatkorpusz ugyanazon mondatában szerepel-e, mint az előző. Ha igen, akkor a költséget ez a tény is csökkenti.

A másik költségfüggvény (*P(n)*) (Prosodic target/prozódiai illeszkedés) definíciója az alábbi:

$$P(n) = \sum_{i=1}^{K+1} v_n(i) * I(u_n(i)) \quad (2),$$

ahol

P(n) a K+1 elem összefűzéséből előálló n-dik alternatíva prozódiai illeszkedési költsége,

$u_n(i)$ az i-dik összefűzött elem az n-dik alternatívában,

$I(u_n(i))$ az i-dik elem prozódiai illeszkedési költsége az ideális prozódiához képest,

$v_n(i)$ az n-dik alternatívában az i-dik és az i+1-dik elem prozódiai illeszkedési költségének súlytényezője.

A prozódiai költség meghatározásánál – az időtengelyi pozíción felül – felhasználjuk az Fo értékének a változását is. Ha nagy Fo ugrás van a két elem között, akkor a költség magas lesz, tehát a két elem nem illeszthető össze.

A végső feladat tehát az

$$X(n) = C(n)+P(n) \quad (3)$$

összeg minimalizálása. A költségfüggvények súlyértékeit iteratív módon, mintegy 500 mondat többszöri szintézisével határoztuk meg. A költségfüggvények alapján először a szószintű, majd a hangszintű optimális elemeket választjuk ki Viterbi-algoritmus [63] segítségével. A költségfüggvény értéke egyben becslést ad a szintetizált mondat minőségére. Ha a költségfüggvény optimalizálás ellenére csak jelentős illesztetlenséget tartalmazó elemeket

találunk a felolvasandó szöveghez, akkor kerül sor a prozódia simítását végző modul alkalmazására. Ez mindenképpen jeltorzulást okoz és gyakran jól hallható a kimeneten. A gyakorlat azt mutatja, hogy ritkán kerül sor e modul alkalmazására.

7. táblázat. Példa a fonetika szabályrendszerből az alacsony költségű vágási pontok kijelölésére. A csatlakozó hangokat a következő szimbólumok jelölik:
C = bármely mássalhangzó; V= bármely magánhangzó, C1= p, t, k, ty, h, f, s, sz, c, cs; C2 = v, j, l, r ; C3= m, n, ny. A hangokat a betűjelükkel adjuk meg.

A megelőző hang a betűjele szerint	A következő, kapcsolódó hang a betűjele szerint	Vágási pont kijelölésének a szabálya	szöveges példa (a csatlakozó hangokat kiemeléssel jelöltük)
V	a) V	a) a hanghatár be van jelölve, ennek ellenére nem célszerű elvágni a hanghatárnál, hanem megfelelő vágási pontot kell keresni visszafelé, vagy előre a hangsorban	*éjszaka**i** **e**sőzésre*
V	b) C	b) a hanghatárnál kell vágni	*nyár**i** **z**áporok*
a) b, d, g, gy	a)V, C2, C3	a) a hanghatáránál kell vágni	*va**d** **v**ihar* *na**gy** **m**eleg*
b) b, d, g, gy	b) önmagával csatlakozik	b) a hosszú hang 70%-ánál kell elvágni, a zárfelpattanás nem lesz benne)	*va**d** **d**örrenés*
c) p, t, k, ty	c) C1 kiv. d)	c) a hanghatáránál kell vágni	*szé**p** **s**ereg*
	d) önmagával csatlakozik	d) a hosszú hang 70%-ánál kell elvágni, a zárfelpattanás nem lesz benne)	*so**k** **k**is*
	e) V, C2, C3	e) a hanghatárnál kell vágni	*szé**p** **f**elhők*
m	a) C kivéve m	a) a hanghatáránál kell vágni	*ne**m** **v**olt*
	b) önmagával	b) a hang 70%-ánál kell elvágni	*ne**m** **m**arad*

Első kísérleti területnek az időjárás-jelentés témakörét választottam. Húsz internetes oldal 2004 áprilisa és 2005 májusa közötti időjárás-jelentéseinek alapján reprezentatív szöveges adatbázist állítottam össze (56.000 mondat, 670.000 szó szintű szövegelem). Ez a szöveges adatbázis túl nagy ahhoz, hogy reális erőforrások mellett (legfeljebb néhány hét alatt) egy professzionális bemondó felolvassa. A méret csökkentésére modellt dolgoztam ki, hogy ne csak az előforduló mintegy 5200 szóalak és a számok jó minőségű felolvasásához szükséges mintegy 230 számelem egy-egy változata kerüljön be a szűkített szöveges adatbázisba, hanem a prozódiai változatosság is megoldott legyen.

Ezeknek a peremfeltételeknek megfelelően a mohó algoritmussal [64] szűkített *szöveges adatbázis*t alakítottam ki. Az eredmény: 5821 mondat, 102.940 szó, ami 488.093 beszédhangnak (fonémának) felel meg.

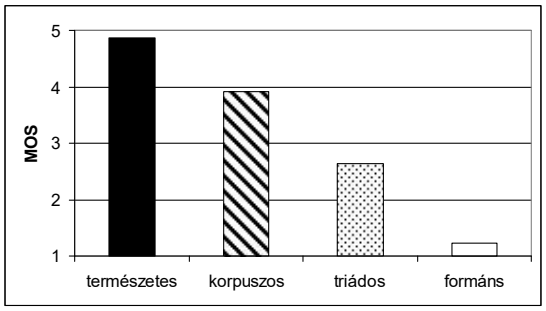
Ezt a szűkített szöveges adatbázist mintegy két hónapos munkával egy professzionális bemondó felolvasásában rögzítettük munkatársaimmal. Ezután utófeldolgozás következett. A hullámformát több szinten címkéztük. A legalsó szinten fonéma (hang) címkékkel történő ellátáshoz félautomatikus eljárást dolgoztam ki a BME TMIT-en fejlesztett beszédfelismerő [37] felhasználásával. A beszédfelismerőt ún. kényszerített üzemmódban (forced alignment, azaz az ismert szövegnek megfelelő hangok pozícióját kellett megjelölni a hullámformában) futtattam. Jelölési módot dolgoztam ki arra az esetre is, ha egy hang csak az eredeti környezetével együtt használható fel, egyedi összefűzésre nem (például az *arra adtam* szókapcsolat közepén az *a_a* egyedi akusztikai jellemzőkkel bír). Ezeket a címkéket, valamint a felhasználásukkal származtatott egyéb adatokat – alapfrekvencia, intenzitás, időtartam és sebesség, stb. – az akusztikai modell legalsó szintjén (hang-szintű összefűzés) használtam fel elsősorban. A következő szinten az egyes hangoknak az adott mondatban, illetve prozódiai egységben elfoglalt pozícióját rögzítettem. A címkék következő (szó) szintjén szintén megjelöltem az egyes szavaknak az adott mondatban, ill. prozódiai egységben elfoglalt pozícióját. Így alakult ki az 2. táblázat szerinti IDO1 beszéd-adatbázis.

Az adatbázis szorosan kötődik a szintézis eljárás 8. ábra (25. o.) szerinti modelljéhez. A fonetikus átíráshoz a ProfiVox rendszer [8] megfelelő modulját alkalmaztam. Az elemválogatás modell kódolását C++ nyelven MSc és PhD hallgatóim (Nagy András, Pesti Péter, ill. Bőhm Tamás, Fék Márk és Zainkó Csaba) végezték.

Számszerű kiértékelés:

A korpusz alapú hullámforma elemválogatás elvén működő gépi szövegfelolvasó rendszer minőségét percepciós teszttel vizsgáltuk [9]. Három magyar rendszert hasonlítottunk össze MOS (Mean Opinion Score) módszerrel, a Multivox formánsszintetizátort, a Profivox diád-triád elemösszefűzéses rendszert és a korpuszos felolvasót. A teszt anyagát a webről kiválasztott időjárás-jelentés 10 mondata alkotta. Ezeket állítottuk elő a fenti rendszerekkel, valamint felolvastattuk őket a beszédkorpusz eredeti bemondójával is. A tesztelőknek tehát összesen 40 mondatot kellett meghallgatni véletlen sorrendben. Minden mondatra egy ítélet született. A tesztet egy interaktív honlap segítségével bonyolítottuk le. A mondatokat 221 személy (egyetemi hallgatók, 185 férfi és 36 nő) hallgatta meg. A teszt elején ismerkedésképpen minden mondattípusból egy-egy mondatot meghallgathattak. A tesztben a mondatokat csak egyszer

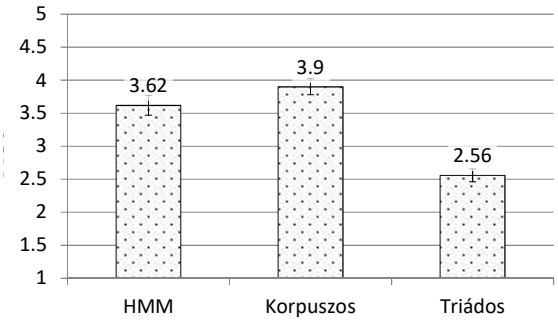
hallották, ismétlésre nem volt mód. A meghallgatások csendes, otthoni környezetben, átlagos (nem professzionális) hangszórókon, illetve fejhallgatókon történtek (a tesztelők a meghallgatás körülményeire vonatkozóan is kitöltöttek adatokat a teszt előtt). A teszt eredményeit a 11. ábra mutatja.



11. ábra. Szubjektív minősítés átlagai a természetes ejtésre és az összehasonlított három szintézis technológiára. [9]

Az értékelésekből látható, hogy a korpusz alapú szintézis hangminősége magasan kiemelkedik a másik két technológiával szemben. A szó, szófűzér alapú összeillesztéssel tehát átléptünk a percepciós megítélésben egy olyan határt, amit a hullámforma összefűzéses (diád/triád alapú) rendszereknél még nem tudtunk elérni, annak ellenére, hogy ott is emberi beszéd részleteit fűztük össze. Feltételezem, hogy a szó képviseli azt a mondatépítő elemet, amelynek szintjén már elégséges egyéni hangjellegzetesség van jelen a hullámformában, hogy a hallgató a beszélő hangszínezetét, egyéni stílusát is felismerje és ennek folytán értékítéletével megközelítse a jó (4) szintet. Természetesen a korpusz alapon szintetizált mondatokban is lehetnek egyenetlenségek a hullámforma folytonosságát illetően (kismértékű dallam ugrások, hangszínezet változások stb.), de úgy tűnik, ha kevés van ezekből, akkor az összegzett ítéletek meghozásakor ezeket a percepciós mechanizmusunk ugyanúgy tűri, feldolgozza, mint az olvasásnál a betűsor feldolgozási mechanizmusunk a betűkimaradásokat, betűhibákat.

Hasonló eredményre jutott egy másik MOS vizsgálat is [19], ahol a korpuszos technológiát a statisztikus parametrikus (HMM, Hidden Markov Model, rejtett Markov-modell) és az elemösszefűzéses triádos technológiával vetettük össze (12. ábra).



12. ábra. Gépi beszéd minőségének összehasonlítása a HMM-TTS, a korpuszos és triád alapú szövegfelolvasó rendszerek között [19].

A HMM és a korpuszos rendszer eltérése nem szignifikáns, míg a HMM és a korpuszos eltérése szignifikáns a triádos megoldástól ($p<0,05$).

Konklúzió:

Kutatásaim eredményeként létrejött az első magyar nyelvű, célorientált korpusz-alapú, hullámforma összefűzést alkalmazó szövegfelolvasó (időjárás-jelentés) rendszer. A rendszer modelljének kialakítása során új módszert dolgoztam ki célorientált szövegfelolvasó szöveges adatbázisának, beszédadatbázisának, prozódiai modelljének és elemkiválasztó algoritmusának létrehozására. Szubjektív tesztekkel megállapítottam, hogy az új megoldás 5-ös skálán mintegy 1,5 ponttal jobb értékelést kap, mint a korábbi legjobb (triádos) rendszer. Bizonyos optimális esetekben ez a módszer az emberi beszédre megtévesztésig hasonló kimenetet képes előállítani. A megoldás helyességét laboratóriumi szubjektív tesztek mellett a metnet.hu internetes tartalom-szolgáltató honlapjába integráltan is sikeresen ellenőriztem. További felhasználását a 9.1 fejezetben ismertetem.

6.2. A korpusz-alapú szövegfelolvasó tématerületekhez történő adaptálása (II.2. tézis)

Egységes eljárást és többszintű modellt dolgoztam ki elsőként a korpusz-alapú hullámforma elemválogatáson alapuló magyar nyelvű szövegfelolvasó technológia különböző tématerületekhez illetve több- vagy kevert nyelvű alkalmazáshoz történő adaptálására. A megoldás működőképességét, valamint az emberi felolvasással való összetéveszthetőségét három (időjárás-jelentés, pályaudvari hangos információ szolgáltatás és árlista-felolvasás) különböző tématerületen igazoltam. Alátámasztó irodalmak: [58], [1]

A korpusz-alapú, hullámforma összefűzést alkalmazó szövegfelolvasó eljárás továbbfejlesztése során megvizsgáltam, hogy milyen feltételek mellett lehet többféle témakörre kiterjeszteni a működést úgy, hogy csak az adatbázist cseréljük ki a rendszerben, a válogatási eljárás algoritmusa pedig ugyanaz. Ez akkor valósítható meg, ha:

- az új tématerülethez rendelkezésre áll elegendő, a prozódiai változatosságot szöveg szinten is biztosító (internetes) forrás és
- a beszéd-adatbázis felolvasására rendelkezésre áll megfelelő idő és bemondó személy.

Azonban előfordulhat, hogy a fenti feltételek nem teljesülnek. Ezért az I.1 tézisben kidolgozott első megoldást adaptálni kellett a jelen tézisben kifejtett két új területre. Arra az esetre, hogy nem érhető el kellő változatosságot biztosító nyilvános, illetve internetes szöveges adatbázis, jó példa a pályaudvari hangos információszolgáltatás, ami *strukturált, de nagy változatosságú témakör.* Ebben az esetben, hazánkban gyakran a mai napig ún. kötött szótáras megoldást alkalmaznak (pl. [65]). Ez azt jelenti, hogy minden menetrend változtatáshoz minden egyes állomáshoz szövegkönyvet kell készíteni és azt minden esetben felolvasva egyedi hangüzenetkészletet kell összeállítani. Minden szöveges üzenethez egyedi (kézzel előírt) összefűzendő hangüzenet készletelem-kombináció tartozik. Ha olyan üzenet merül fel, ami nincs benne az előre tervezett készletben, akkor a tájékoztatást csak az adott helyen egy dolgozó közvetlen bemondásával lehet megoldani.

Erre a célorientált témakörre *új, többszintű modell*t dolgoztam ki. Mondatsémákat alakítottam ki, melyekhez változó tömbtípusokat rendeltem. A mondatséma egy egész üzenet leírására ad példát a tömbök elemeinek a felhasználásával. Az alábbiakban példát adok tömbtípusokra:

Vonatnevek
Pl. *Füzér IC* vagy *Füzér InterCity*
Magyar állomás nevek
Pl. *Érd*
Országon kívüli állomás nevek
Pl. *Bari, München*
Változó elem nélküli mondatok

pl. *Kérjük, a vágány mellett vigyázzanak.*
Változó elem nélküli részmondatok
Pl. *Felhívjuk tisztelt utasaink figyelmét….*
Kifejezések, 2-3 szóból álló összetartozó szövegrészek
pl. ..... *közlekedő személykocsik…..*

A mondatsémák és a tömbtípusok felhasználásával kellő prozódiai változatosságú szöveges adatbázis kialakítására alkalmas algoritmust dolgoztam ki. Ezt automatikus szoftver eljárással valósítottam meg, melynek segítségével előállt a felolvasandó szövegadatbázis. Ennek felolvasásával és az II.1 tézisben leírtak szerinti feldolgozásával állt elő a megfelelő beszéd-adatbázis. A fonetikai átíró célszerű módosításával és a beszédadatbázis cseréjével a II.1 tézisben bemutatott időjárás-felolvasó rendszer alkalmassá vált pályaudvari hangos információ-szolgáltatásra. Noha a közel emberi minőség csak a szöveges mondatsémáknak megfelelő üzenetekre garantálható, a rendszer alkalmas tetszőleges szöveg érthető felolvasására is. A fentiek szerinti modellt és módszert alkalmaztam vasúti pályaudvari mintarendszer létrehozására (2. táblázat szerinti **PALYA1** beszéd-adatbázis). A mintarendszer 2009 óta az egyik észak-magyarországi pályaudvaron működött. 2014 óta pedig több mint száz MÁV állomáson és megállóhelyen vezették be megoldásunkat (részletesen ld. 9.1 fejezet). A modell és a módszer alkalmazható VOLÁN, BKV és repülőtéri hangos utastájékoztató rendszer kialakítására is.

A harmadik vizsgált célorientált tématerület esetén tervezett szöveges adatbázis nem, vagy csak korlátozottan állt rendelkezésre, viszont elérhető volt egy távközlési mobilszolgáltató ügyfélszolgálati rendszere automatizált bemondásainak szöveg- és hanganyaga mintegy három évre visszamenően (2. táblázat szerinti **UGYF1** beszéd-adatbázis nyersanyaga, első lépésben 3747 mondat, 69057 szó). Ekkor alapvetően a meglevő adatbázisból lehetett építkezni, kiegészítő hangfelvételekre minimális mértékben volt lehetőség.

A kísérletek során először azt vizsgáltam, hogy az adott peremfeltételek mellett lehetséges-e a cél-terület követelményeinek megfelelő rendszert létrehozni. Ez minden korábbinál nagyobb elvárást jelentett, hiszen hagyományosan ezeket az üzeneteket mindig emberek olvasták be, és a felhasználók hozzászoktak ehhez. Másrészt a mobilszolgáltatókkal szemben az ügyfelek nagyobb minőségi elvárásokat támasztanak, mint a korábbi cél-területek (időjárás-jelentés és közlekedési információ) szolgáltatóival szemben. Ezért a felhasználónak *nem szabad, hogy feltűnjön, hogy gép olvassa fel* az az üzeneteket. Első lépésként a rendelkezésre álló szöveg- és beszéd-adatbázisokat egymáshoz illesztettük. Ellenőriztük a korábban már említett kényszerített felismerés módszerével, hogy pontosan a szövegnek felel-e meg a hangfelvétel és hibás esetben kézi javítást végeztünk, melynek eredményeként kialakult az UGYF1 beszédadatbázis. Az

adatbázist illesztettük a korpusz-alapú infrastruktúrába. Majd kísérleteket végeztünk a szöveges mintához hasonlító, de azzal nem azonos mondatokkal. Noha a kimenet érthető volt, azonban megakadások, prozódiai hibák előfordultak benne. Mivel nem volt lehetőség az adatbázis lényeges bővítésére, így a minőség javítására csak a tématerület szűkítése merülhetett fel. A lehetséges alternatívák közül a viszonylag szűkebb, de nagy gyakorlati jelentőséggel bíró területként az **árlista-felolvasás**t választottam ki.

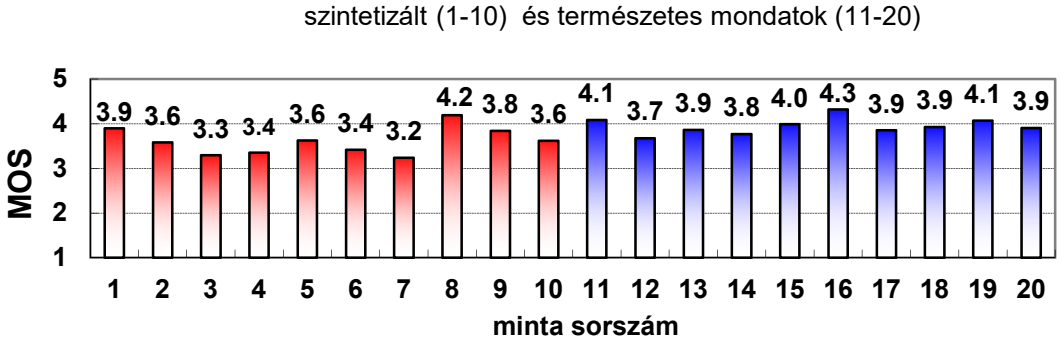
A célterület tehát egy mobil távközlési szolgáltató árlistájának (készülékek különböző előfizetési típusokhoz – feltöltős, flotta, egyéni, stb. – tartozó ára, szolgáltatások díjai, fizetési feltételek, stb.) az emberi felolvasástól az átlagos felhasználó számára nem feltűnően elütő gépi felolvasása volt. Ebben az esetben egy *új, hibrid, félautomata megoldás*t dolgoztam ki. Mivel mind a készülékek, mind a szolgáltatások esetében gyakran jelennek meg új, előre nehezen, vagy nem tervezhető nevek (pl. iPhone) az adatbázis bővítését könnyen lehetővé tevő eljárást dolgoztam ki. Első lépésként az UGYF1 adatbázis tematikus szűkítésével létrehoztam az **ARU1** beszédadatbázist. Ez tartalmazta az adott területre érvényes termékek és szolgáltatások megfelelő prozódiai lefedettségét biztosító szöveges- és hangmintákat. Ezt véletlenszerű bemondások szubjektív tesztelésével ellenőriztem. Ha új termék vagy szolgáltatás jelenik meg, annak a kiejtését a rendszer kezelője ellenőrizheti. Ha a megoldás nem megfelelő, akkor az új elemről előre meghatározott ún. vivő mondat(ok)ba illesztve hangfelvételt készít és mind a szöveges, mind a hangos formát eljuttatja a fejlesztőkhöz. A magyar kiejtéstől eltérő szöveges formákat kivételszótárban feleltetjük meg a fonetikus leírásnak, majd automatikusan, kényszerített felismerés módszerével generáljuk a megfelelő címkéket és az új hangmintát hozzáadjuk az adatbázishoz. Így az adatbázis folyamatosan bővíthető a piacon és a nyelvben megjelenő új termékekkel. Ez az új, hibrid megoldás már megfelelő minőséget biztosított.

A modell kódolását C++ nyelven MSc és PhD hallgatóim (Bartalis Mátyás, ill. Kiss Géza, Tóth Bálint és Zainkó Csaba) végezték.

Számszerű kiértékelés:

Az időjárás-felolvasás minőségét az I.1-es tézisben ismertetett módon, valamint a metnet.hu honlapba integráltan ellenőriztem. Hasonló módon és eredménnyel került sor a pályaudvari hangos utastájékoztató laboratóriumi ellenőrzésére is. 2009 óta pedig a rendszer az egyik észak-magyarországi MÁV állomáson működik felhasználói panasz nélkül. A legkomolyabb felhasználói követelmény (és valószínűleg a legtöbb felhasználó) az árlista-felolvasó megoldással kapcsolatban merült fel, ezért annak vizsgálatát részletesebben ismertetem.

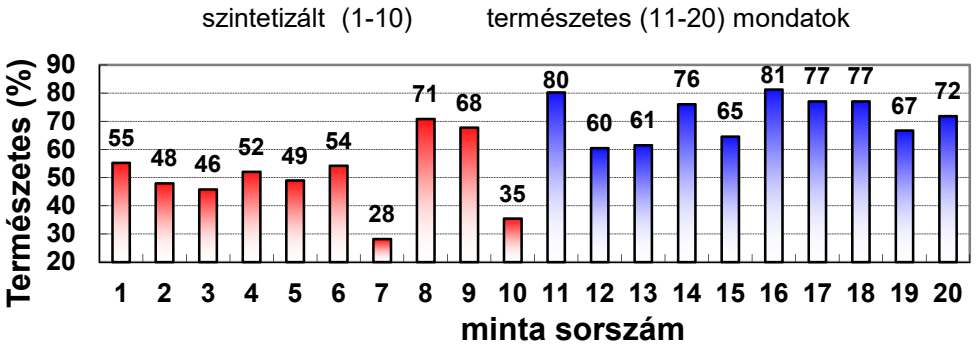
Web-alapú szubjektív tesztet végeztünk, 93 (67 férfi, 26 nő, átlagosan 32 éves) magyar anyanyelvű, ismert halláskárosodással nem bíró teszt alannyal. Átlagosan 38 percig tartott a vizsgálat. A hangmintákat szabványos RTE-LTP GSM kódolóval állítottuk elő, hogy a teszten a hangminőség hasonló legyen, mintha az ügyfelek a vállalatot mobiltelefonon hívnák. A teszt első részében 5 pontos MOS skálán értékeltek 10 szintetizált (1-10, 13. ábra) és 10 természetes bemondást (11-20, 13. ábra). Mind a 20 minta eltérő tartalmú volt.



13. ábra. Az árlista mondatok átlagos szubjektív minősége.

Az eredmények szerint a természetes bemondások jobb értékelést kaptak (átlag: 3,95, szórás: 0,18, 3,7 – 4,3 között) míg a szintetizált bemondások valamivel alacsonyabb, de jónak mondható értéket értek el (átlag: 3,60 szórás: 0,3, 3,2 – 4,2 között). Noha szignifikáns a különbség a két változat között, ez csak 0,3 pont. Ez kevesebb, mint a fele a szabványos PCM (4,3) és a GSMRPE-LTP (3,5) MOS értéke különbségének. Megjegyzendő, hogy a távközlési szabványok MOS számítása sokkal összetettebb folyamat, mint amit a mi lehetőségeink megengedtek. Minden szintetizált változat elérte, vagy meghaladta a 3,2 értéket. Mindkét minta 930-930 bemondást tartalmazott, tehát az eredmények meglehetősen megbízhatók.

A teszt második részében azt vizsgáltuk, hogy egy mondatot természetesnek vagy szintetizáltnak értékelnek a tesztalanyok. 10-10 szintetizált (1-10, 14. ábra) és természetes (11-20, 14. ábra) mintát hallgattak meg véletlen sorrendben. A 14. ábra jelzi a "természetes" értékelések arányát. Az 50%-os érték felel meg a véletlenszerű eloszlásnak. Egymintás T-tesztekkel értékeltük, hogy hány minta különbözik a véletlen értékeléstől jelentősen. Minden természetes bemondást (11-20) szignifikáns mértékben természetesnek értékeltek ($p=0,05$) míg a szintetizált mondatok (1-10) közül kettőt (8, 9) értékeltek szignifikáns mértékben „természetes" kategóriába, kettőt (7, 10) pedig szignifikáns mértékben a „szintetizált" kategóriába osztályoztak.



14. ábra. A teszt mintákra adott "természetes" értékelések aránya.

A többi esetben nem lehetett szignifikanciát megállapítani (azaz a teszt alanyok nem tudták egyértelműen eldönteni, hogy a szintetizált minta természetes vagy szintetizált forrásból származik-e). Ez nem jelenti azt, hogy a szintetizált változatok ugyanolyan jók, mint a természetesek. Csak azt állíthatjuk, hogy a tesztalanyok nem tudták megkülönböztetni, tehát ha kimondottan nem figyelnek rá, valószínűleg nem tűnik fel a különbség. A GSM kódoló fontos szerepet játszhat ebben. 2009-2015 között ez a rendszer is éles üzemben működött egy mobil távközlési szolgáltatónál.

Konklúzió:

Kutatásaim során modellt és eljárást dolgoztam ki három eltérő jellegű célorientált tématerület korpusz-alapú hullámforma elemösszefűzéses gépi szövegfelolvasó technológiájának kidolgozására. Igazoltam, hogy a korpusz-alapú technológia mindhárom célterületen képes az emberi beszéddel összetéveszthető gépi beszédet előállítani

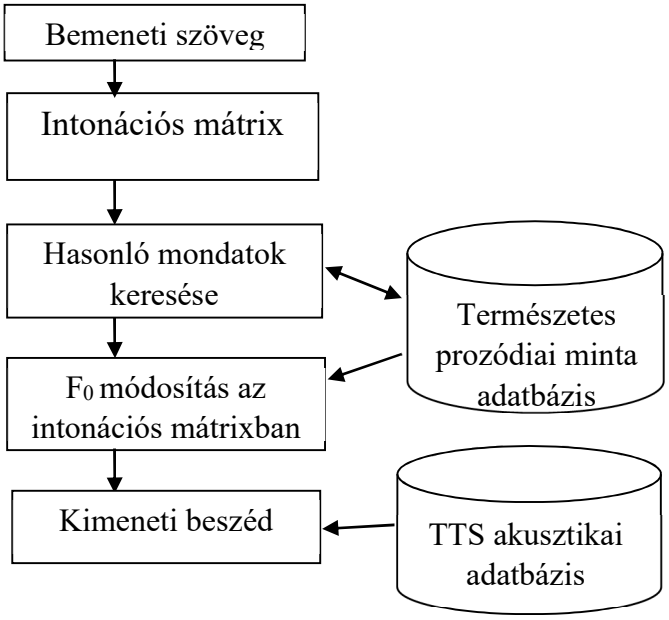
6.3. A gépi szövegfelolvasás prozódiai változatosságának megvalósítása (II.3. tézis)

Új módszert dolgoztam ki prozódiai frázisok hasonlósága alapján képzett prozódiai csoportok létrehozásához és ezekből nem determinisztikus válogatással gépi szövegfelolvasó rendszerek prozódiai változatosságát tettem lehetővé. Megmutattam, hogy egy magyar nyelvű megvalósítás során a felhasználók ezt a módszert a hagyományos szabály-alapú és a II.1-es tézis szerinti indirekt megoldásnál is jobbnak értékelték. Ez a prozódiai modell alkalmazható a hagyományos elemösszefűzéses, a korpusz-alapú és a HMM rendszerekben egyaránt.

Alátámasztó irodalmak: [66], [67], [68]

A gépi szövegfelolvasás elfogadásának egyik korlátja az, hogy a rendszerek döntő többsége determinisztikus működésű, azonos szöveg-bemenetre mindig azonos hullámforma-kimenetet produkál. Rövid szövegek esetén ez kevésbé zavaró, viszont ismétlődő mondatszerkezetű, hosszabb szövegek nehezen elviselhetően, monotonnak hangzanak. A természetes kiejtésben még egyszerű mondatok is (pl. Jó napot kívánok) kis részleteikben jelentősen változnak egyazon

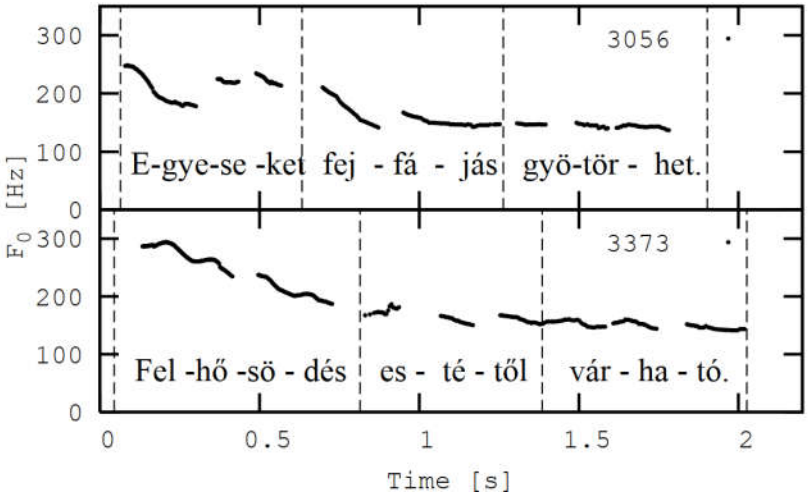
személy ismétlései során. [69] azt állapította meg 1000 Mandarin nyelvű mondat kétszeri ismétlésén, hogy azonos mondat esetén az alapfrekvencia és a szótag időtartam is jelentősen változhat. Új módszeremet a 15. ábra alapján ismertetem.



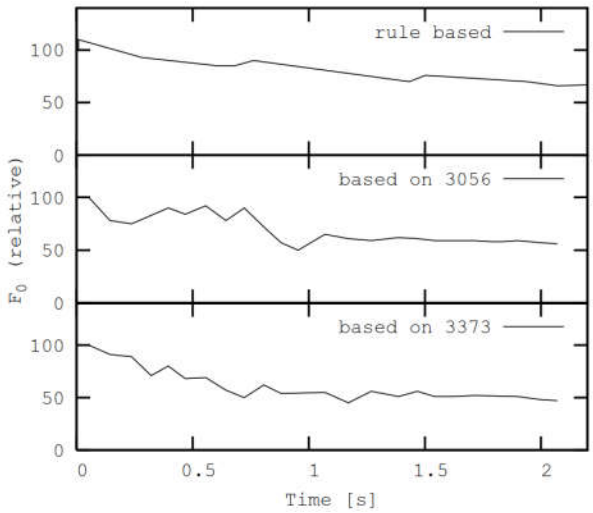
15. ábra A javasolt prozódia meghatározási módszer áttekintése [66] *alapján*

Lényege az, hogy nagyméretű beszédadatbázisból prozódiai frázis (elemi gondolati egység, célszerűen két akusztikai szünet közti hullámforma elem) szintű mintákat gyűjtök magas szintű szintaktikai/fonológiai jegyek alapján (szószám, szótagszám, a közlés modalitása, stb.). Kihasználom, hogy a magyar nyelvben a szóhangsúly jellemzően az első szótagon van. Például azonos struktúrájú mintákat szolgáltatnak a „ Jó reggelt kívánok!”, „ Jó napot kívánok!”, „ Jó estét kívánok!” közlések. Ezen mintákból **prozódiaiminta-adatbázis**t építek. Angol nyelven a hasonlóság meghatározása megoldható, de lényegesen bonyolultabb, többek között a változó pozíciójú szóhangsúly miatt.

A 16. ábra két teljesen eltérő szövegű, de hasonló szótag szerkezetű mondat alapfrekvencia menetét mutatja. Látszik, hogy a két intonációs jelleg nagyon hasonló, de nem azonos. Az eltérő időtartamot normalizálással kompenzálom.



16. ábra. Két hasonló szótagszerkezetű mintamondat alapfrekvencia menete. A függőleges szaggatott vonal a szóhatárokat jelzi ([66]



17. ábra. A 3056-os sorszámú mondat három intonációs alternatívája [66]

A szintetizálandó szöveget prozódiai frázisokra bontom és minden prozódiai frázisra megvizsgálom, hogy a prozódiai minta adatbázisban vannak-e illeszkedő (azonos vagy hasonló) minták szöveg szinten. Ha vannak, akkor véletlenszerűen választok közülük. Ily módon garantálható, hogy ugyanannak a szövegnek a többszöri, ismétlődő szintézise esetén változatos, de az adott szöveg köznapi ejtésének megfelelő szintetizált kimenethez jutunk. A módszert magyar nyelvű mintarendszerben valósítottam meg és teszteltem.

A megoldást a 17. ábra segítségével mutatom be. A 3056-os sorszámú mondat (Egyeseket fejfájás gyötörhet., ld. 16. ábra) szintézisének szabályalapú intonációja a felső részen látható. A középső része jelzi a természetes bemondásból származtatott alapfrekvencia menetet. Az ábra

alján pedig a 16. ábra szerinti 3373-as mondatból származtatott kontúr követhető. Mindhárom megoldás megfelel a magyar nyelv szabályainak, viszont jól azonosíthatóan eltér egymástól.

A tématerületen a tézisben ismertetett eredmények felhasználásával, a tudás alapú megközelítés és a gépi tanulás módszereinek kombinálásával továbbra is végzünk kutatásokat [70].

A modell kódolását C++ nyelven MSc és PhD hallgatóim (Csapó Tamás és Zainkó Csaba) végezték.

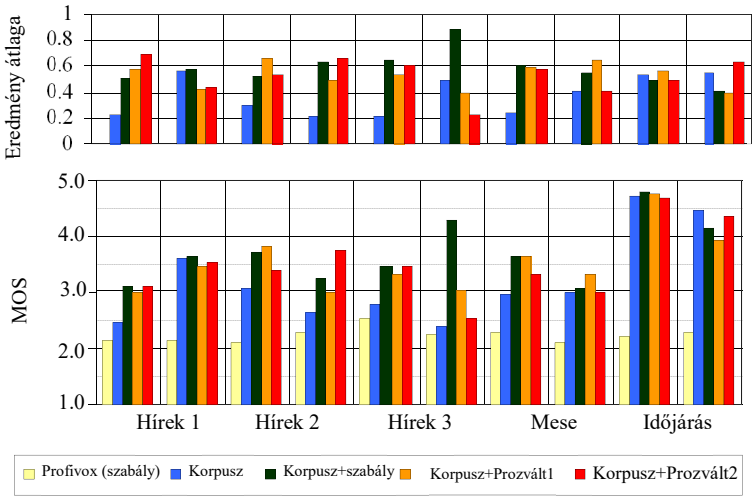
Számszerű kiértékelés:

Különböző célterületekhez tartozó mondatokat vizsgáltunk a módszer felhasználásával. Az IDO1, FON1, PALYA1, SZAM és RADIO beszédadatbázisok felhasználásával származtattam a prozódiai mintákat. Az I.1 tézis szerinti rendszerbe integráltam az új módszer prozódiai megoldását és így folytattam le a kiértékelését. Öt témakörben gyűjtöttünk szövegmintákat (a jelölésében Tömegközlekedés=Hírek 1, Gazdasági=Hírek 2, Sport=Hírek 3, Mese és Időjárás). Minden témakörben két mondatot választottunk ki. Minden mondatból öt változatot szintetizáltunk. Az elsőt a ProfiVox rendszer triád-alapú változatával, ami szabály-alapú prozódiai modellt alkalmaz. A másodikat a II.1 tézis szerinti rendszer pozíció-alapú indirekt prozódiai modelljével (korpusz). A harmadik a korpusz-alapú rendszer a ProfiVox rendszer szabály-alapú prozódiai modelljét alkalmazza célfüggvényként (korpusz+szabály). A negyedik változat egy „pontos” prozódiai séma hasonlóság alapján választ mintát (korpusz+Prozvált1), az ötödik pedig „lazább” prozódiai hasonlósági mérték alapján működik (korpusz+Prozvált2). Ha nincs a prozódiai mintakészletben hasonló minta, akkor a negyedik és ötödik esetben a prozódiai célfüggvény azonos a harmadik változattal.

Az értékelést két módszerrel végeztük. Minden mondat természetességére vonatkozóan MOS értékelést folytattunk le 5*10=50 mondatra. A rendszerek közötti kisebb különbségek felderítése érdekében pedig páros összehasonlítással CMOS tesztet alkalmaztunk. Mivel előzetes vizsgálataink szerint a triád-alapú rendszer minősége lényegesen rosszabb, ezért a 10 mondat 4 változatát hasonlítottuk össze.

A tesztben 93 (67 férfi és 26 nő, átlagosan 32 éves) ép hallású, magyar anyanyelvű személy vett részt az I.2 tézis szerinti rendszerek minősítésével egy időben. A 18. ábra alsó része mutatja mind a 10 mondat MOS értékeit. Minden csoport első oszlopa mutatja a triád-alapú TTS értékeit (2.1-2.5). Ezek megegyeznek a II.1 tézis értékelésekor kapott eredményekkel (11. ábra, 32.o., 12. ábra, 33.o.). A korpusz-alapú rendszerek teszteredménye két csoportba sorolható. Az első 8 csoport kevésbé természetes, mint az utolsó kettő. Ez várható is, hiszen az utolsó két rendszer

szöveges bemenete egyezett meg a célorientált rendszer adatbázisának területével. Az Időjárás témakörben az új módszer ezért nem eredményezett javulást. A másik három esetben azonban az új módszer egyértelmű előrelépést jelentett a I.1 tézis szerinti alaprendszerhez képest.



18. ábra. A 10 tesztmondat páros összehasonlítású (CMOS felső) és természetesség (MOS, alsó) minősítése.

A 18. ábra felső része mutatja a páros összehasonlítás eredményeit. A természetesebbnek tartott változat kapott 1,0, a kevésbé természetes 0,0 értéket. Ha a tesztelő a két változatot egyformának értékelte, akkor mindkettő 0,5-et kapott. Az ábra az átlagokat mutatja. A Tukey-HSD post hoc teszt szignifikáns különbségeket jelzett. Az új módszer valamennyi változatát szignifikánsan természetesebbnek értékelték ($p < 0,05$) az I.1. tézis szerinti alaprendszernél.

Konklúzió:

Az általam kidolgozott adatbázis alapú prozódiai modellezés módszere magyar nyelvű mintarendszeren a korábbi megoldásoknál szignifikánsan jobbnak bizonyult (18. ábra).

7. Statisztikus parametrikus gépi szövegfelolvasó rendszerek (III. téziscsoport)

A statisztikus parametrikus gépi szövegfelolvasás elsőként a rejtett Markov modell elméletére alapozva jött létre. A Markov modell matematikai kereteit már a XX. század elején lefektették [71]. Az IBM-nél Fred Jelinek és kutatócsoportja dolgozta ki ezen elmélet alapján az első gépi beszédfelismerő rendszert a 70-es években [72]. Ennek alapján jöttek létre az első, kereskedelemben kapható, nagyszótárú beszédfelismerő rendszerek (IBM Tangora, Dragon Systems, Philips dictation, stb.). A beszédfelismerésben elért sikerek vezettek oda, hogy felmerült az elmélet alkalmazása gépi szövegfelolvasás céljaira is. Az első ilyen rendszert a nagoyai egyetemen Tokuda professzor irányításával fejlesztették ki [39].

Felismertem, hogy a statisztikus parametrikus gépi felolvasó rendszer optimális megoldást jelenthet beszédsérült emberek rehabilitációjának támogatásához (ld. IV.3 tézis). Kezdeményeztem egy rejtett Markov modell (HMM) alapú magyar nyelvű gépi szövegfelolvasó (TTS) rendszer létrehozását és meghatároztam a modellalkotás lépéseit. A modell létrehozása folyamán meghatároztam a tanításhoz szükséges beszédadatbázis szerkezetét, koncepciót és modellt alkottam a statisztikus parametrikus modellhez illeszkedő beszédkódolók létrehozásához, valamint módszert dolgoztam ki rövid és kérdő mondatok jobb minőségű szintéziséhez.

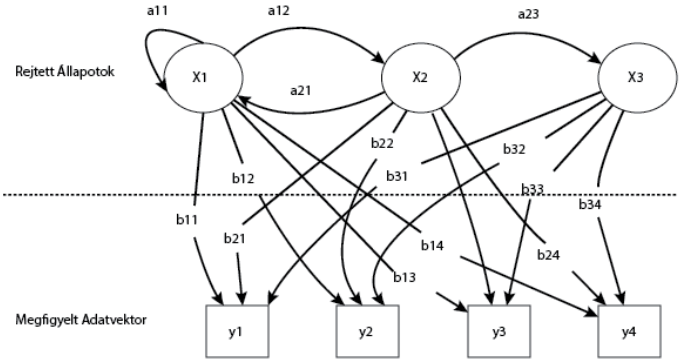
7.1. A rejtett Markov modell alapú magyar nyelvű gépi felolvasó rendszer (III.1 Tézis)

Azonosítottam az újonnan megalkotandó vagy adaptálandó rendszermodulokat az első gépi tanuláson alapuló magyar nyelvű gépi szövegfelolvasó rendszer kialakításához. Létrehoztam egy olyan adatstruktúra-modellt, ami alapján az ezen az elven alapuló gépi szövegfelolvasó rendszer hatékonyan megvalósítható. Alátámasztó irodalmak: [73], [51]

Felismertem, hogy a HMM megközelítés új és jelentős kutatási irány, ezért figyelemmel kísértem fejlődését. Ahogy látható volt az elsősorban japán és angol nyelven folytatott kísérletekről szóló publikációkból, hogy a rendszer stabilizálódott és nyilvános forráskódban is elérhető volt egy változata, kezdeményeztem a magyar nyelvre alkalmazott megoldás kidolgozását. A megközelítés előnyei közé tartozik az, hogy a megfelelő modellek és módszertan megalkotása után nagyrészt automatikusan lehet egy adott személy hangjára emlékeztető gépi szövegfelolvasó rendszert létrehozni. Ezzel lehetővé válhat ún. hangbankok (voice bank) létrehozása, ahol az egészséges személy hangját tárolják és szükség esetén annak alapján

beszédkommunikációs segédeszközök számára az illető személy hangjára emlékeztető gépi hangot tudnak előállítani. Ma már ilyen kereskedelmi rendszerek is léteznek (pl. https://www.vocalid.co).

A 19. ábra szemlélteti a HMM modell alapgondolatát.



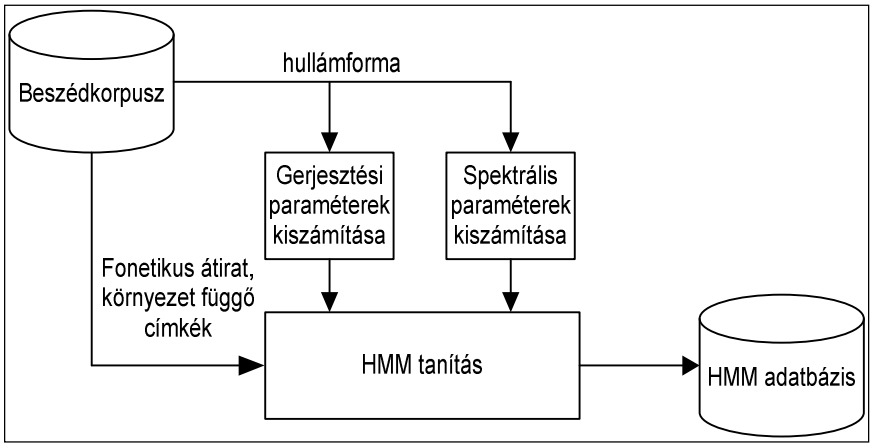
19. ábra. A HMM modell alapgondolata

Az y1, y2, … adatvektorokat tudjuk megfigyelni. Ezeket az *x1, x2,* … állapotok közti átmenetek során emittálja a modell. Az állapotokat nem ismerjük, ezért „rejtett” a modell. Az állapotátmenetek valószínűségét adják meg az *a11, a12, a21*, … súlyok. A *b11, b12, b13, b14* valószínűségek azt jelzik, hogy az adott állapotban milyen valószínűséggel bocsátja ki a modell a megfelelő adatvektort.

Tehát a feladat az, hogy az adatvektorok ismeretében becsüljük meg, hogy milyen állapotátmenet-sorozat valósult meg a modellben. Beszédfelismerés esetén valamilyen lényegkiemelt paraméter (pl. cepstrum együtthatók) az adatvektorok, az állapotok pedig a kimondott beszédhangoknak felelnek meg. Beszédszintézis során pedig az adatvektorokat a bemeneti szöveg (vagy az ahhoz tartozó intonációs mátrix adatai) jelentik, az állapotoknak pedig egy beszédkódoló paramétervektorai (egyebek között spektrális adatok) felelnek meg. A tanítás (training) folyamata során egy referencia adatbázis alapján becsüljük meg a modell adott állapotteréhez tartozó (*xi* állapotok) *aij* és *bkl* valószínűségeket. A szintézis folyamata során pedig a beadott adatvektorokhoz tartozó vezérlő paramétereket generálja a rendszer a beszédkódoló számára.

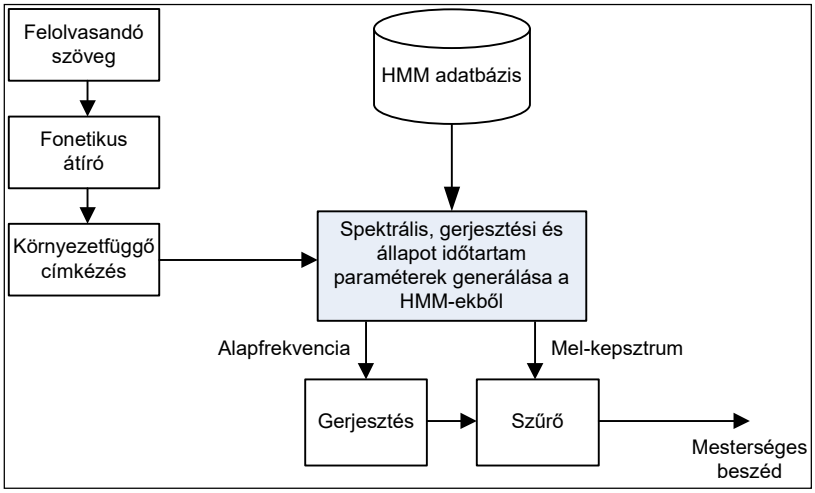
A HMM beszédszintetizátor által használt tanító rendszer felépítését a 20. ábra mutatja. A kutatás első lépéseként a gerjesztési paraméterek és a spektrális paraméterek tekintetében a HTS alaprendszer jellemzőit használtuk fel. Csak a beszédkorpusz, a fonetikus átirat és a környezetfüggő címkék előállítását végző modulok képezik az önálló eredményeket.

A III.2 tézisben ismertetem a beszédkódolásra vonatkozó kutatási eredményeimet, melyek idővel természetesen a tanítási és a generálási eljárásokban is megjelentek.



20. ábra. A HMM alapú szövegfelolvasó tanítása [74]

A szintézis folyamat során a felolvasandó szövegből először előállítjuk a felolvasandó szöveghez tartozó hangsorozatot és az úgynevezett környezetfüggő címkéket (hangok pozíciója a szóban, szavaké a mondatban, hangsúlyok, stb.). Ezekkel tanítottuk be a betanítási fázisban a HMM-eket. A generálás során a beszédkódoló spektrális, gerjesztési és időtartam paraméterei állnak elő a betanítás során kialakított HMM modellekből álló adatbázis segítségével. A paraméterek tartalmazzák a beszédhangok spektrális jellemzőit és egyidejűleg a prozódiát megvalósító adatokat is. A HMM alapú beszédszintetizátor felépítését a 21. ábra mutatja.



21. ábra. A HMM alapú szövegfelolvasó felépítése [74]

A tanító adatbázisnak ebben az esetben a lehető legnagyobb mértékben le kell fednie a magyar nyelv beszédhangállományát. Kiinduló szövegállományként Vicsi Klára és munkatársai

beszédfelismerési célokra kialakított szöveglistáját használtam fel [75]. Ezt a beszédszintézis céljaihoz szükséges elemekkel, például rövid és kérdő mondatokkal egészítettem ki. Ennél a módszernél jelentős az eltérés a hullámforma elemösszefűzéses eljáráshoz képest, ugyanis itt egy-egy hangkapcsolatból minél több realizációt kell megvalósítani annak érdekében, hogy a statisztikus modell változatos szövegbemenethez is megfelelő beszédkódoló vezérlő paramétereket generáljon. Prozódiai szempontból fontos, hogy amíg az elemösszefűzéses esetben egyenletes hangmagassággal (természetellenesen) kell a diád-triád adatbázishoz a felolvasást elvégezni, addig az elemkiválasztásos (korpuszos) megoldásnál célszerű korlátozott dinamikát alkalmazni, hogy az elemek között ne legyenek jelentős paraméterugrások. A statisztikus parametrikus (HMM) eljárás során pedig a természetes prozódiájú felolvasás alkalmazható a hangfelvétel készítése során, ugyanis az eljárás simítási tulajdonsággal rendelkezik. Megterveztem a fenti elveknek megfelelően a beszédadatbázis szerkezetét és munkatársaimmal félautomatikus eljárást dolgoztam ki az adatbázisok hatékony elkészítéséhez [51]. Ezek a kutatási eredmények alapozták meg kutatócsoportunk további tevékenységét a témakörben, amely eddig egyebek mellett Tóth Bálint [19] és Csapó Tamás [18] PhD fokozatához vezetett.

Számszerű kiértékelés

A tézisben létrehozott új megoldások önmagukban nem mérhetőek, azonban a 12. ábra (33. o.) szerinti rendszerszintű értékelés kimutatta, hogy a HMM alapú megoldás ugyan rosszabb MOS értékeket kapott, mint az elemkiválasztásos (korpuszos) alternatíva, azonban ez a különbség nem szignifikáns. A megoldás azonban jobban automatizálható, kisebb fejlesztési és futtatási erőforrásigénnyel bír.

Konklúzió

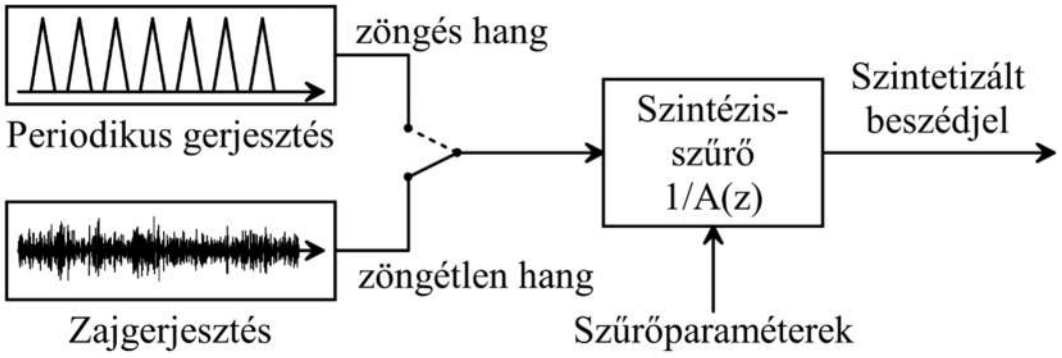
A tézisben bemutatott eredmények alapján alakult ki az első magyar nyelvű HMM TTS rendszer.

7.2. A HMM TTS rendszer minőségének javítása (III.2 Tézis)

Új elven, a maradékjelre alkalmazott elemkiválasztásos eljáráson alapuló, megvalósítást elősegítő koncepciót és modellt alkottam a HMM TTS rendszerben alkalmazandó jobb minőségű beszédkódolók létrehozásához. Alátámasztó irodalom: [76]

A beszéd kódolásának egyszerűsített forrásszűrő modellje a 22. ábra látható. A modell megvalósításával néhány paraméter (zöngés/zöngétlen kapcsoló, a zönge alapfrekvenciája, a jel intenzitása, szűrő paraméterek) megfelelő időközönkénti (tipikusan 5-30 ms) beállításával érthető

beszédjel állítható elő egy adott paraméterhalmaz feldolgozásával. Formánszintézis esetén a szűrő paraméterek formáns frekvencia és sávszélesség értékek. Ezek meghatározása azonban nehezen automatizálható, ezért az első HMM TTS rendszerekben a forrásszűrőt lineáris predikciós módszerek segítségével határozták meg. A modell egyszerűsége egyszerre előny és hátrány. Előnye, hogy könnyen és hatékonyan megvalósítható, így alkalmas kis erőforrású eszközökben (pl. mobiltelefonok) történő alkalmazásra. Hátránya viszont, hogy elvi okokból nem alkalmas kevert gerjesztésű hangok (pl. *'z, zs'*) helyes akusztikájú létrehozására. Igényesebb minőséget lehet elérni a STRAIGHT [77] eljárással, azonban ez egyrészt bonyolult, jelentős erőforrások mellett lehet futtatni, ezért nem alkalmas az egyszerű, széles körben használt mobileszközökön történő felhasználásra. Másrészt gyakorlati szempontból lényeges, hogy csak kutatási célokra szabad felhasználású a módszer, kereskedelmi célokra bonyolult jogdíjfizetési előírások vonatkoznak rá.



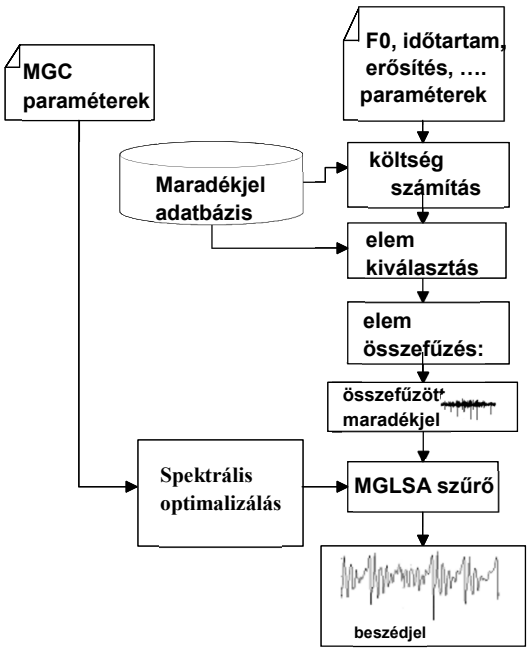
22. ábra. A beszédkódolás egyszerűsített modellje [1]

Célom a tézis szerinti kutatási témában az, hogy átlagos mobileszközön történő futtatásra alkalmas, de a STRAIGHT eljárás minőségét elérő, beszédszintézis feladatokra optimalizált, statisztikus parametrikus beszédszintézisre jól alkalmazható kódoló eljárás jöjjön létre. A kutatás alapgondolata az, hogy a forrásszűrő modellen belül a szűrő modellezése az eddigi megoldásokkal megfelelő eredményt hoz, azonban a forrás modellezésében sok megoldásra váró feladat van. Még a legkorszerűbb parametrikus rendszereknél is előfordul a zöngés hangoknál hallható háttér zümmögés (buzziness) és többnyire gondot jelent az elvi modellekkel nehezen leírható, de a valós beszédben gyakran előforduló ún. rekedtes, glottalizált hangok (creaky voice) reprodukálása is.

Felismertem, hogy a II. téziscsoport szerinti elemkiválasztásos rendszer alapgondolata alkalmazható a forrásszűrő lineáris predikciós (LPC) modelljében a beszédjel ún. inverz szűrése

eredményeként létrejövő maradékjel (residual signal) felhasználásával létrehozott adatbázisból megfelelő algoritmus segítségével kiválasztott elemek forrásjelként való felhasználására.

A kódoló alapgondolata a 23. ábra alapján követhető. A spektrális jellemzésre általában a Mel-Generalized Cepstral elemzési (MGC) [78] módszert használtuk. A forrásjelet egy célszerűen az adott beszélő személy hangfelvételeiből előállított maradékjel-adatbázis elemeinek összefűzéséből állítjuk elő.



23. ábra. A beszédkódolás maradékjel elemkiválasztáson alapuló modellje [79]

A kiválasztás egy költségfüggvény alapján történik, melynek az alapvető prozódiai paramétereken túl (zöngés/zöngétlen, ill. kevert gerjesztés jelleg, alapfrekvencia, időtartam, intenzitás) olyan kiegészítő paraméterek lehetnek a bemeneti adatai, mint a zöngésség, ill. a zajosság foka, a glottalizáció és annak foka, stb.

A tézis szerinti eredmények további kutatási irányokra vezettek [18], [79], [80], [81], [82], [83], [84].

Számszerű kiértékelés

A tézisben létrehozott új megoldások az előző bekezdésben felsorolt, a továbbfejlesztését megvalósító rendszereket bemutató publikációk szerint mind lényegesen jobb minőséget értek el, mint a klasszikus periodikus vagy zajforrást alkalmazó beszédkódoló algoritmus.

Konklúzió

A tézisben bemutatott eredmények alapján jelentős mértékben továbbfejlesztésre került a magyar nyelvű HMM TTS rendszer.

7.3. Rövid és kérdő mondatok jobb minőségű megvalósítása (III.3. Tézis.)

Kidolgoztam a magyar kérdő mondatok alapfrekvencia-idő függvényeinek statisztikai modellezését gépi beszéd-előállításhoz. Alátámasztó irodalmak: [85], [86]

A statisztikus parametrikus rendszerek jellemzője, hogy a beszédadatbázisukban található elemekhez optimalizálják a modelljeiket. Általában a hétköznapi közlésekben a kijelentő mondatok vannak túlsúlyban, ezért ezek fordulnak elő a legnagyobb számban az adatbázisokban és következésképpen a modellek ezeket tudják a legjobb minőségben előállítani. Kutatásaink során felfigyeltem arra, hogy a rövid (1-3 szótagos), valamint a kérdő mondatok jelentős részének a prozódiai megvalósítása nem kielégítő hangzást eredményez. Az első vizsgálatok során kiderült, hogy ilyen típusú mondatok kis számban jelentek meg az adatbázisban. Ezért kutatást folytattunk arra vonatkozóan, hogyan lehet ezt a kérdést megoldani. Kiderült, hogy ha túl sok ilyen mondatot viszünk be az adatbázisba, akkor a gyakoribb (átlagos és hosszabb) kijelentő mondatok minősége romlik.

Első eredményes megoldásunk az volt, hogy nagyszámú, különböző nemű, korú, beszédstílusú személytől származó, kisebb-nagyobb méretű adatbázist összevontunk, az így kialakított nagyméretű adatbázisból tanítottunk egy ún. átlaghangot (average voice), majd azt az átlaghangot adaptáltuk az adott feladathoz illeszkedő kisebb adatbázissal [85]. Ily módon a magyar nyelvre általánosan jellemző tulajdonságok is bekerültek a statisztikus modellbe, de lehetővé vált az adott személyhez vagy alkalmazáshoz kötődő jellemzők megfelelő beillesztése is a modellbe. A megoldás javította a megcélzott mondattípusok szintézisét anélkül, hogy a nagyobb gyakorisággal generált kijelentő mondatok minőségét rontotta volna. Azonban az eldöntendő kérdő mondatokat így is gyakran kijelentő mondatnak észlelték a teszalanyok.

Ezt a megoldást továbbfejlesztve eljárást dolgoztunk ki minden kérdő mondat típus hatékonyabb szintézisére. Az alapgondolat az volt, hogy az I. tézis szerinti rendszerben [8] alkalmazott, a kérdő mondatok valamennyi típusát jól észlelhetően generálni képes modellt [87] kombináljuk a HMM technológia emberihez hasonlóbb minőséget biztosító előnyével. Ezt kétféle módon modelleztük [86].

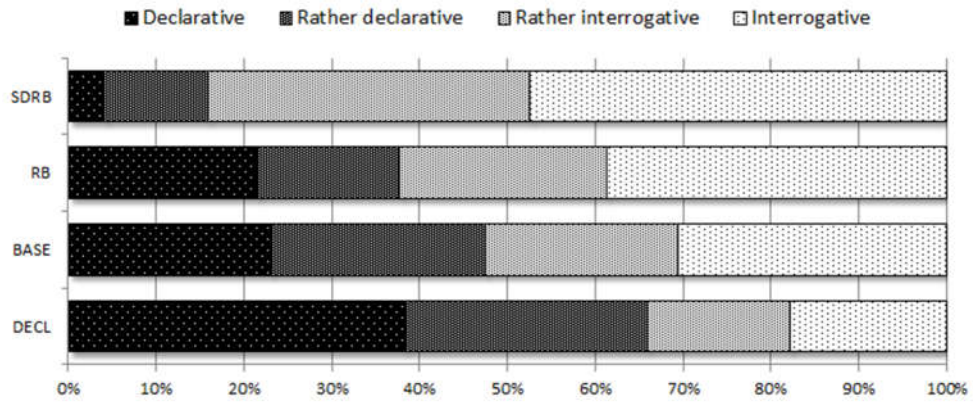
Az első esetben az I. téziscsoport szerinti szabályalapú modell szerinti hangsúlyos, ill. hangsúlytalan szótagok alapfrekvencia adatait vetjük össze idővetemítés után a HMM modell által javasoltakkal. Amennyiben ütközést találunk, a szabályalapú modell hangsúly előírásai szerint átskálázzuk a paramétereket. Azaz, ha hangsúlyt észlelünk olyan helyen, ahol nem lehet, akkor az alapfrekvencia értékét csökkentjük. Ha pedig nincs hangsúly ott, ahol szükséges, akkor növeljük az alapfrekvenciát. Az eljárás alkalmazásához a HMM modell betanításához szükséges

beszédadatbázisból kiszámítottuk a beszélő alapfrekvencia gyakorisági eloszlását. A továbbiakban ezt az eljárást RB-vel (szabályalapú, Rule Based) jelöljük.

A második módszer az első módszer továbbfejlesztéseként a normalizálást nem csak két feltételhez köti. A leggyakoribb kérdő mondat típusokhoz egy az I. tézis szerinti modellben alkalmazott ún. kulcspontokat (key points) rendel. A beszédadatbázisból pedig kiszámoljuk a megfelelő mondat ípusok megfelelő kulcspontjaihoz tartozó, a beszélő személytől függő átlagértékeket. Majd ezt normalizáljuk az adott beszélő átlagos alapfrekvenciájára. A skálázást pedig dinamikusan, a kulcspontok közti lineáris interpoláció segítségével végezzük el. Ezt a módszert a továbbiakban SDRB-vel (személyfüggő, szabályalapú, Speaker Dependent Rule Based) jelöljük. A kutatást Nagy Péter doktorandusszal és Tóth Bálinttal közösen végeztük.

Számszerű kiértékelés

A tézisben ismertetett koncepció alapján kialakított rendszerek kiértékelésének eredményét mutatja a 24. ábra. A BASE jelölés a közvetlenül a beszédadatbázisból betanított prozódiát generáló rendszerre vonatkozik. A DECL jelölésű (kijelentő alapú, DECLarative) rendszer esetében a tesztben alkalmazott kérdő mondatokat kijelentő mondatként szintetizáltuk a BASE rendszerrel. A DECL rendszer kivételével minden, a bemenetre adott szintetizált mondat szövege kérdőjellel végződött. A teszt szöveglistája valamennyi kérdő mondattípust tartalmazta. A teszthez 30 kérdő mondatot választottunk ki. A tesztalanyok mind a négy rendszerből 7-7 mondatot értékeltek. A feladatuk az volt, hogy egy négyelemű skálán értékeljék a mondat modalitását: Kijelentő, Inkább kijelentő, Inkább kérdő, Kérdő. Természetesen nem tudták, hogy mikor van kérdőjel a bemeneti szöveg végén.



24. ábra. Kérdő mondatok szintézisére szolgáló alternatívák kiértékelése [86]

A kísérletben 38 személy (12 nő és 26 férfi) vett részt. Mindannyian magyar anyanyelvűek és ép hallásúak. A teszteket webes felületen keresztül végeztük. Az átlagos életkor 46 év, a legfiatalabb személy 24, a legidősebb 83 éves volt. Egyikük sem beszédtechnológiai szakértő.

Az eredményekből megállapítható, hogy a legjobb rendszer az SDRB, azonban az RB változat is lényegesen jobb, mint a HMM alapváltozat (BASE). A DECL változat mutatja a szemantikus információ prozódiai észlelést befolyásoló hatását.

Konklúzió

A tézisben bemutatott eredmények alapján a magyar nyelvű HMM TTS rendszer a korábbiaknál lényegesen jobb minőségű rövid kijelentő és kérdő mondatok előállítására vált képessé.

8. Multimodális beszédinformációs rendszerek (IV. téziscsoport)

A gépi beszédkeltés és beszédfelismerés technológiáit hosszú időn keresztül elsősorban telefonvonalon keresztül folyó ember-gép interakciókban alkalmazták. A 2000-es évek elejétől viszont egyre nagyobb jelentőségre tesznek szert a grafikus és a beszéd felhasználói felületeket (esetleg más, pl. taktilis, gesztus eljárásokat is) kombináló megoldások. Ezen a kutatási területen új megoldást dolgoztam ki a modalitások szinkronizált kezelésére (IV.1 tézis) és azt alkalmaztam egy speciális kommunikációs segédeszköz kidolgozására (IV.3 tézis). Kidolgoztam egy hatékony, gyors akusztikus üzenetforma – a spemoticon – elméletét és megvalósításának módszertanát (IV.2. tézis).

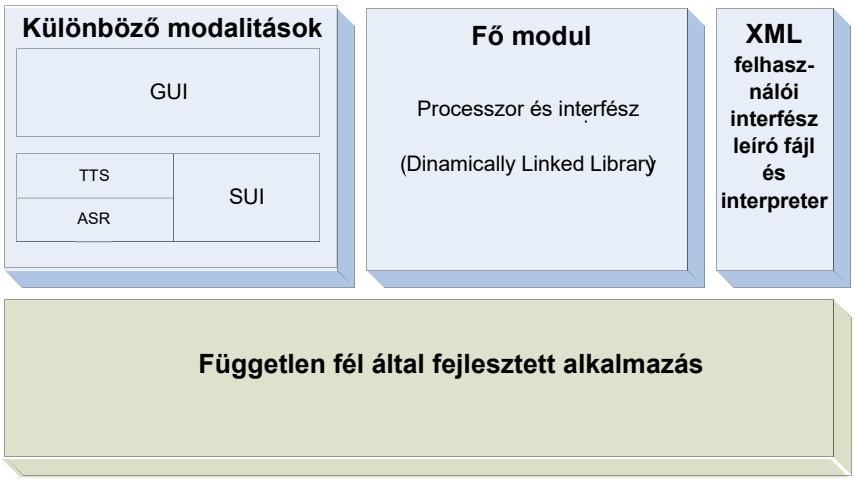
8.1. Mobil felhasználói felületek modalitásainak szinkronizálása (IV.1. tézis)

Új, skálázható, multimodális leíró nyelvet alkalmazó eljárást dolgoztam ki mobil multimodális felhasználói felületek modalitásainak szinkronizálására. A módszer működőképességét a grafikus és a beszéd modalitás szinkronizálását megvalósító mintaalkalmazásokkal igazoltam.
Alátámasztó irodalmak: [88], [89], [90]

Az információs rendszerekben hagyományosan unimodális felhasználói felületeket alkalmaznak, a számítógépes környezetben jellemzően grafikusakat (Graphical User Interface, GUI). A beszédtechnológiát pedig jellemzően telefonos beszédinformációs rendszerekben alkalmazzák, tehát ez beszéd-alapú felhasználói felület (Speech User Interface, SUI vagy Voice User Interface, VUI). Ergonómiai megközelítéssel korábban vizsgálták, hogy adott feladatot milyen bemeneti és kimeneti modalitásokkal lehet a leghatékonyabban megoldani [91]. Megállapították, hogy a feladattól függően más-más ki- és bemeneti modalitások az optimálisak. Magas szintű megközelítést is alkalmaztak multimodális dialógusok hatékony tervezésére [92]. Speciális leíró nyelvet is terveztek erre a célra [93].

A GUI és a SUI megoldásokhoz is készültek formalizált leírások, azonban ezek nagyon változatos, a különböző mobil platformok és a modalitások közti egyidejű átjárást nem támogató megoldások, az átjárást általában a különböző platformok GUI változatai között támogatják.

Ezért olyan formalizált eljárás kidolgozását tűztem ki célul, ami egyszerű, széleskörűen használt leíró nyelven alapul és támogatja legalább a GUI és SUI modalitások szinkronizált használatát. A megoldás alapja a minden mobil platformon alkalmazott XML leíró nyelvet támogató middleware megoldás, amellyel minimális programozással, csupán a felületleíró fájlok elkészítésével lehetséges működő alkalmazások készítése. Az eljárás blokkdiagramját a 25. ábra mutatja.



25. ábra. A skálázható, multimodális leíró nyelvet támogató architektúra.

Az alapötlet az, hogy XML alapú leíró nyelvet és egy azt feldolgozó modult hozunk létre a támogatott platformokon. Az így támogatott GUI és SUI felületeket független (3rd party) fejlesztők vastag kliens alkalmazások fejlesztésénél is felhasználhatják.

A '*Fő* modul' kapja meg a felhasználói interfész leírását az '*XML felhasználói leíró fájl és interpreter*' elemek felhasználásával. A '*Különböző modalitások*' modul kezeli egységes koncepciók szerint a GUI és a SUI szinkronizált megvalósítását. A szövegfelolvasó (TTS) és beszédfelismerő (ASR) nyilvános interfészen keresztül kapcsolódnak a 'Fő modul' blokkhoz, így az adott platformon elérhető bármilyen beszédtechnológiai elem alkalmazható. Az XML leíró fájlra a 26. ábra ad egyszerű példát.

```
<UserControl Name=”myTextBox” Type=”textbox”
ServiceClass=”Temperature” Size=”120px” Posx=”10” Posy=”5”
Input=”keys” Input=”voice” Output=”GUI” Output=”SUI”
Action=”setTemperature” Security=”All”>Please define the in-
car temperature</UserControl>
```

26. ábra. XML felhasználói interfész leíró fájl (User Interface Description File).

A példa szerint egy 120 pixel széles szövegdoboz (textbox) kerül a képernyő (10, 5) pozíciójába. Ez a hőmérséklet (Temperature) szolgáltatás osztályba tartozik. Értékét a mobilkészülék billentyűjével és beszéddel egyaránt beállíthatjuk. A biztonsági előírások (Security Policy) szerint az értéket a felhasználó bármikor módosíthatja. A beállított érték mind beszéddel, mind szövegesen kijelzésre kerül. Ha a beállított hőmérséklet érték változik, a setTemperature függvény kerül meghívásra. Az adatkéréskor a *„Kérem, adja meg a jármű kívánt belső hőmérsékletét”* (Please define the in-car temperature) üzenet kerül szövegesen és hangban is kijelzésre.

Az új módszert különböző Windows és Symbian okostelefonos platformokon és mintaalkalmazásokban MSc és PhD hallgatóim implementálták (Decsi István, Hámor Tamás, Doan Thi Bich Huyen, ill. Tóth Bálint).

Számszerű kiértékelés:

Az új módszert mintaalkalmazások fejlesztése kapcsán teszteltem. Az eljárás hatékonyságát egy grafikus és beszéd-felülettel egyaránt rendelkező RSS-felolvasó hagyományos programkódolással és az új módszerrel készített változatán mutatom be [94]. A jellemző paramétereket a 8. táblázat mutatja.

8. táblázat. Multimodális RSS-felolvasó fejlesztésének jellemző paraméterei hagyományos és az új módszerrel [94]

Mért adat	Első verzió	Második verzió
SUI leírások mérete	58 + 1274 = **1332 sor** (37,4 kbyte)	**261 sor** (10,3 kbyte)
Forrás sorainak száma (kommentek nélkül)	919 sor	815 sor
Program indulása az első felület felépüléséig	8 sec	19-21 sec
Hírcímek letöltésének és megjelenítésének ideje	60 sec	8-9 sec
Részletes híroldalak letöltésének és megjelenítésének ideje	2-3 sec	8-9 sec
Program indulása, RSS letöltés és egy hírfelolvasás elkezdése	**70 sec**	**40 sec**

A sorok abszolút értékei kevésbé érdekesek, a fontosak az arányok. Az új módszerrel gyorsabb működést és rövidebb programkódot (ezzel gyorsabb fejlesztést) lehet elérni.

Konklúzió:

Az általam kidolgozott eljárás alkalmas felhasználói felületek modalitásainak szinkronizálására. Az eljárás működőképességét és hatékonyságát mintaalkalmazásokkal igazoltam.

8.2. Kommunikációs kontextust jelző akusztikus jelkészlet előállítása (IV.2. tézis)

Kidolgoztam kommunikációs kontextust jelző új akusztikus jelkészlet (spemoticon-ok) elméletét és modelljét valamint annak megvalósítási módszerét gépi szövegfelolvasó eszközrendszerére alapozva. Megalkottam egyfajta jelkészlet csoportot. Objektív paraméter beállítások módszerével és szubjektív tesztekkel igazoltam a módszer eredményességét.
Alátámasztó irodalom: [95]

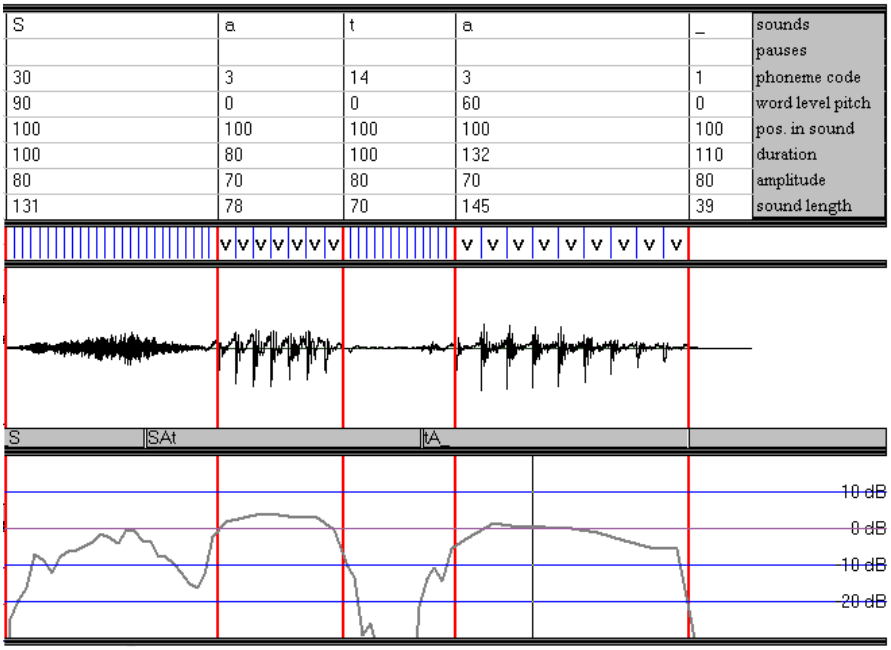
A közismert hangjelzések alkalmazása szituációk kifejezésére általánosan használt technika a mindennapi életben (sziréna, villamoscsengő, duda, makogás, nevetés stb.). Az ilyen hangjelzéseket *akusztikus ikon*oknak (acoustic icons) nevezik [96]. Az „*earcon*”-ok olyan mesterségesen tervezett és keltett hangjelek [97], melyekről meg kell tanulnunk, hogy mit fejeznek ki egy adott élethelyzetben (például frekvencia modulált jel frekvenciája arányos egy alakzat magasságával). Az akusztikus ikonok előnye az, hogy mivel jól ismert hangjelenségek, ezért könnyen lehet újabb fogalmi asszociációt rendelni hozzájuk (pl. fájl törlést WC öblítés hangjához). Viszont az earcon-okkal ellentétben nehéz objektív adatokkal, ill. paraméterekkel jellemezni és módosítani őket. A két megközelítés érdekes összekapcsolási kísérlete a *spearcon* (speech earcon) és a *spindex* (speech index) [98]. A spearcon olyan (akár géppel keltett) beszéd, amit az érthetőség határáig felgyorsítanak, így gyorsan értelmezhető és megjegyezhető leírását adhatja egy (vagy néhány) szónak. A spindex egy listában keresést azzal segíti, hogy a lista elemeinek néhány hangját (vagy szavát) gyorsítja fel, majd – ha a felhasználó nem lép tovább – rövid szünet után átlagos sebességgel olvassa fel a lista elemét. A szövegfelolvasó minél nagyobb felgyorsításának lehetőségét a vakügyi célokat figyelembe véve korábban, a ProfiVox rendszer [8] tervezésekor kutatótársaimmal figyelembe vettük.

Kutatási célomat Jeon és Walker [98] tevékenységével párhuzamosan, attól függetlenül, ahhoz hasonlóan, de általánosabban tűztem ki. Olyan modellt kívántam létrehozni, mellyel az ember számára ismerős akusztikus jelen alapuló olyan elemkészlet alakítható ki, mely objektív paraméterekkel jellemezhető.

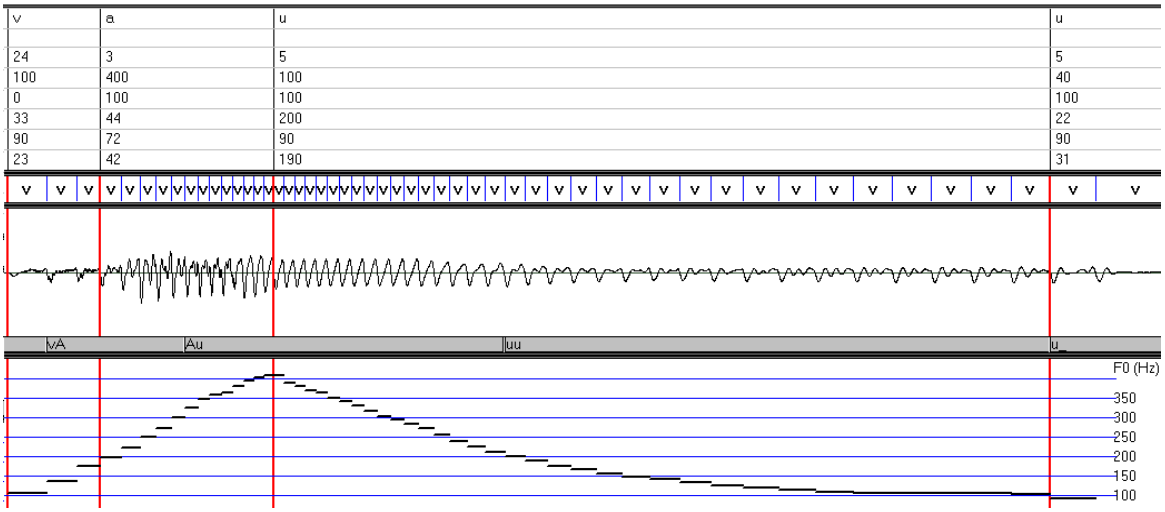
A *spemoticon* (speech-based emoticon) modelljének alapgondolata az, hogy az egyrészt emberei beszédből kivágott elemi jelhalmazból – kísérleteinkben a ProfiVox gépi szövegfelolvasó akusztikai adatbázisából –, mint kiindulási forrásból épül fel. Másrészt viszont egy megfelelő célszoftver (fejlesztői rendszer) segítségével annak alapvető paramétereit (alapfrekvencia, időtartam, intenzitás) jól kézben tartott eljárással megváltoztatjuk és nem beszéd jellegű, hanem adott kommunikációs kontextusra jellemző, jól azonosítható hangminták készletét állítjuk elő.

A kísérleteket a ProfiVox rendszer fejlesztői rendszere (27. ábra) segítségével végeztem. Egy spemoticon szerkesztésekor tetszőleges (akár értelmes, akár értelmetlen, ugyanis a módszernek

nincs elvi okokból nyelvfüggése) hangsort írunk az első sorban található szövegmezőbe, majd a szintetizátorral létrehozzuk annak adatbázis-alapú hullámformáját. Ezután az adatmátrix paramétereinek változtatásával beállítjuk azt a hangzást, amit az adott kommunikációs kontextushoz szükségesnek ítélünk. Ez a munkafázis kreatív szabad alkotás. Az eredményt meg tudjuk hallgatni. Az iteratív folyamatot segíti a hullámforma és az intenzitás (vagy az alapfrekvencia változás, ill. a spektrum) kijelzése.



27. ábra. A szerkesztői felület felépítése: az első sorban egy értelmetlen két szótagú szöveg látható, alatta az objektív prozódiai adatokat tartalmazó adatmátrix, majd középen a hullámforma és alul az intenzitás görbe.



28. ábra. A vau szövegből kiinduló, módosított hangjelenség (spemoticon jelölt).

A folyamat végeredményeként előálló hangjelenségre (spemoticon jelöltre) mutat példát a 28. ábra, a *vau* szövegből kiindulva. Az 'u' időtartama jelentősen megnyújtásra került és az 'a' és az 'u' átmenetére egy alapfrekvencia csúcs került, ami a hang végére cseng le.

Ezzel a módszerrel gyorsan és rugalmasan állíthatók elő hangjelenségek (spemoticon jelöltek). Ezek közül azokat tekintjük majd spemoticonnak, melyeket a megcélzott felhasználói közösség döntő része egyértelműen azonosítani tud egy szituációval.

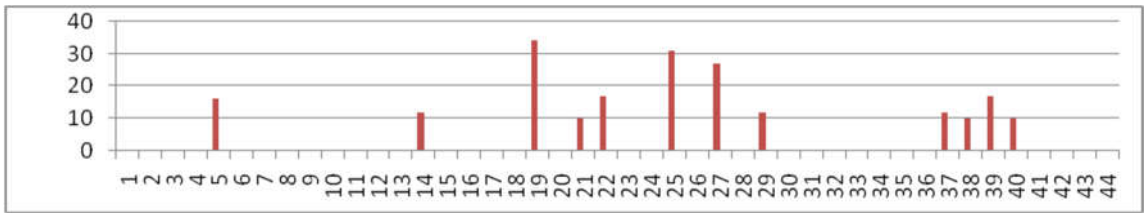
A modell kidolgozásában segítségemre volt Olaszy Gábor. A teszt kialakításában közreműködött Csapó Tamás PhD hallgatóm.

Számszerű kiértékelés:

Az eljárás minősítéséhez az alábbi hét érzelmi és kontextuális kategóriát határoztuk meg (a kategóriák a felhasználási célnak megfelelően változhatnak):

1. Folytasd, tetszik. Kérem, ismételd meg! (pozitív érzelem)
2. Élvezem, rendben van, mehet. (pozitív érzelem)
3. Ez nem jó, nekem. Ne csináld! Zavar engem! (negatív akció)
4. Dühös vagyok! Utállak! Ez zavar! (konfliktus, negatív helyzet)
5. Szomorú vagyok! Nincs jókedven. (rosszkedv és következményei)
6. Vigyázat! Figyelj! (figyelmeztetés, aggódás)
7. Gratulálok! Ez siker volt! (pozitív értékelés, dicséret)

A tesztalanyoknak (54 ép hallású személy) a fenti hét kategória közül kellett egyet társítani a hallott hangjelenséghez egy webes felületen keresztül. Átlagosan 15 perc alatt végeztek a véletlenszerűen lejátszott 44 minta értékelésével úgy, hogy egy-egy mintát annyiszor hallgattak meg, ahányszor akarták. Az értékelés során megvizsgáltuk, hogy melyek azok a hangjelenségek, melyeket a tesztalanyok többsége egy adott kategóriába sorolt. Feltételeztük, hogy a nem egyértelmű hangjelenségek véletlenszerűen szóródnak a kategóriák között. A 44 mintából így 9 spemoticon alakult ki (az 1., 2. és 7. pozitív kategóriákra öt, a 4. és 5. negatív kategóriákra négy). Egy válasz eloszlás mintát mutat a 29. ábra (a függőleges tengely válasz gyakoriságot jelez).



29. ábra. Az 1 pozitív kategóriához (Folytasd, tetszik. Kérem, ismételd meg!) tartozó válaszok eloszlása. Spemoticon hangminták: 19, 25 és 27.

Konklúzió:

Az általam kidolgozott eljárás alkalmazása igazolta, hogy a gépi beszédkeltési technológián alapuló modell felhasználható kommunikációs kontextust (és érzelmet) kifejező akusztikus jelkészlet (spemoticonok) létrehozásához.

8.3. Multimodális felhasználói felületek beszédsérült emberek támogatására (IV.3. tézis)

Új módszert dolgoztam ki multimodális felhasználói felületek hatékony felhasználására beszédsérült emberek kommunikációjának támogatására. A módszert a gépi szövegfelolvasó rendszerekben többféle szövegbeviteli formára és eszközplatformra (asztali számítógép, notebook, okostelefon, tablet) alkalmaztam. Alátámasztó irodalmak: [99], [100]

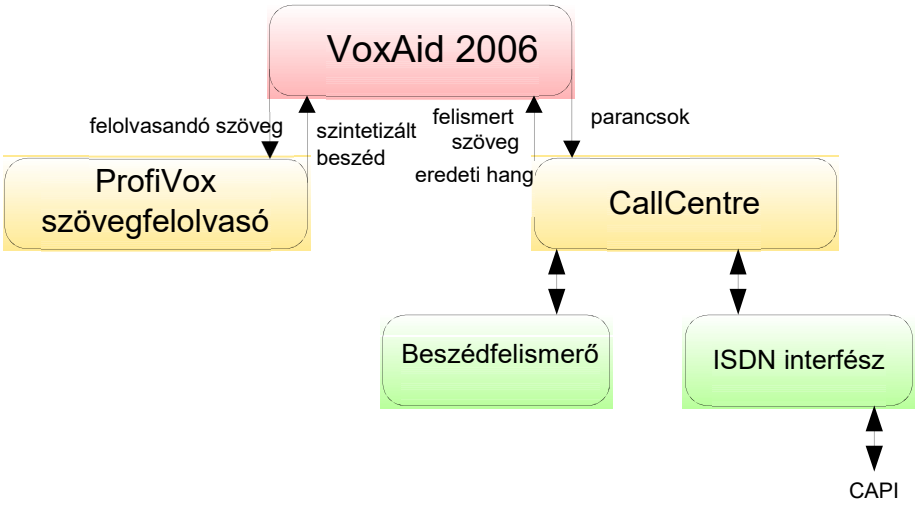
A multimodális felhasználói felületek egyik első alkalmazási területe a rehabilitáció volt. Már Bánó Miklós [22] szabadalmi leírásában is szerepel az, hogy a „világtalan gépíró hallgassa a leírt szöveget, a néma gépíró pedig a berendezés segélyével hangosan közölhesse mondani valóját”. Korábbi kutatásaink során [101] már kidolgoztuk egy multimodális (kép+beszéd) kommunikációs segédeszköz prototípusát beszédsérült, de ép hallású személyek számára. Ez a rendszer (laptop) a negyvenes évei során elszenvedett stroke után haláláig, mintegy 15 éven keresztül biztosította egy beszédképességét elvesztett hölgy kommunikációját. Az ebben a tézisben ismertetett kutatásaink során ezt a koncepciót illesztettem az elmúlt 20 év technológiai fejlődése által biztosított lehetőségekhez.

A beszédsérült emberek kommunikációjának egy jelentős korlátja az, hogy nem tudnak telefonálni. A VoxAid (másik márkaneve StrokeAid) rendszer első változatával ezt beszédképességüket elvesztett, de ép hallású személyek számára tettük lehetővé [101]. Kutatásaim következő lépésében azt tűztem ki célul, hogy siketnéma személyek számára is lehetővé váljon a kapcsolt távközlő hálózaton folytatott beszélgetés ép hallású és beszédű személyekkel, teljesen automatikus eljárás segítségével.

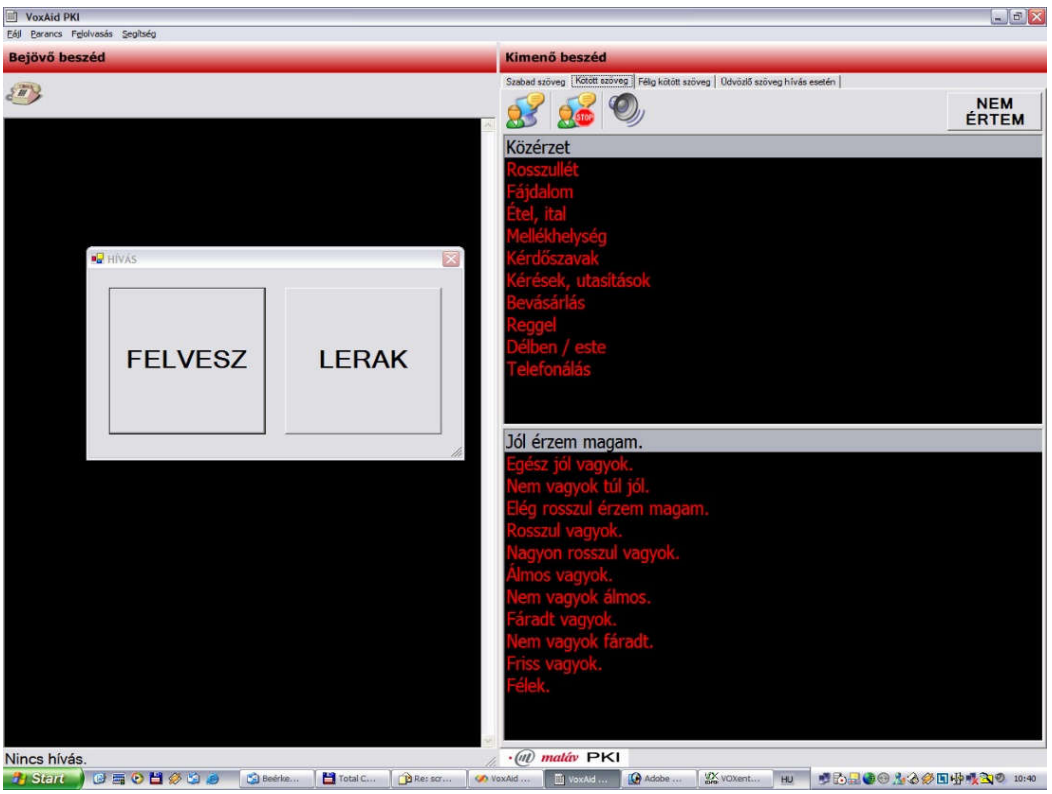
A megoldás azon alapul, hogy akkoriban (2004-5) vált elérhetővé magyar nyelven is PC-s környezetben nagy szótárú, telefonvonalon keresztül is működő beszédfelismerési technológia [102]. Korábban a magyarhoz hasonló szerkezetű finn nyelv esetében megállapították, hogy hallássérült emberek viszonylag nagy (akár 20%-ot is elérő) fonéma szintű hibát is fel tudnak dolgozni dialógus szituációban [103].

Ezért azt a hipotézist tettem fel, hogy van esély arra, hogy a partner beszédét beszédfelismerő dolgozza fel és adja meg szöveges formában, míg számára a választ a gépi szövegfelolvasó szintetizált beszédje adja meg. A siketnéma személy pedig a számítógép megosztott képernyőjén keresztül láthatja egyrészt a számára küldött üzenetet, másrészt pedig oda írhatja be a válasznak

szánt mondandót. A megoldás szoftverkomponenseinek a kapcsolatát a 30. ábra mutatja be. A telefonvonalat (ISDN kapcsolat) szabványos CAPI interfészen keresztül érte el a rendszer. A beszédfelismerő és a telefonvonal integrált kezelését megvalósító CallCentre modul lehetővé teszi a telefonvonal ki- és bemeneti hangcsatornájának elérését a VoxAid2006 alkalmazás számára. Ennek az a célja, hogy ha ép hallású személy van a közelben vagy a felhasználó csak beszéd-, ill. hallássérült, akkor lehetőség van az ép szerv használatára.



30. ábra. A VoxAid 2006 alkalmazás és a kapcsolódó szoftver komponensek.



31. ábra. A VoxAid2006 rendszer felhasználói felülete telefonhívás során.

A 31. ábra illusztrálja a felhasználói képernyőt. Ha hívás érkezik, villogó ikon jelzi. Ha „felvesszük a kagylót”, akkor egy előre beállítható rendszerüzenet hangzik el. Így lehet elkerülni, hogy a gépi hangot a hívó viccnek gondolja. A hívó fél a beszédfelismerő szótárát is lekérdezheti nyomógombok segítségével az SMS betűválasztásához hasonló módszerrel. A felhasználói felület számos egyéb interakció-optimalizáló megoldást is tartalmaz, melyek ismertetése túlmegy a jelen dolgozat keretein.

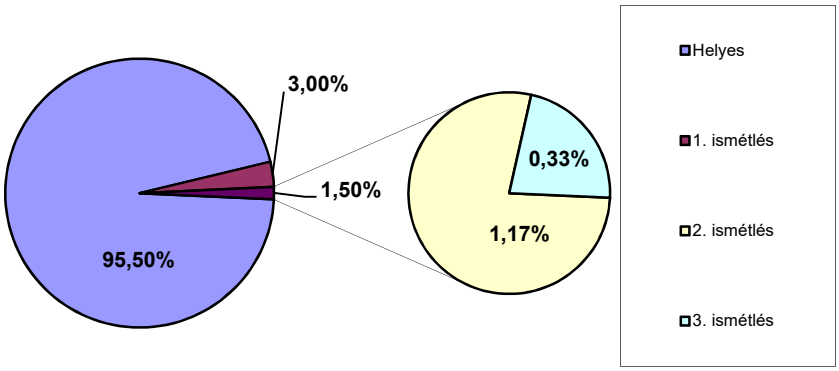
A kutatás következő lépésében azt vizsgáltam, hogyan lehet a számítógépekre már jól kidolgozott koncepciót átvinni okostelefonokra is. Az első kísérleteket Symbian és Windows Phone operációs rendszeren végeztem ép hallású, beszédsérült emberek támogatására. Megállapítottam, hogy az okostelefonok kis kijelzőjén nehézkes az üzenetkategóriák és az üzenetek szerkesztése, ezért kidolgoztam az asztali számítógépes változat és az okostelefonos kiegészítő (Android és Windows Phone 6.5 operációs rendszer alatt futó) integrált változatát. A nagy képernyős rendszeren szerkesztett üzenetstruktúrát változtatás nélkül át tudja venni az okostelefonos változat.

A rendszer fejlesztése iteratív, felhasználó-orientált megközelítéssel történik. Felhasználók a beszédsérült személyek mellett a rehabilitációban közreműködő logopédusok is. Az ő kérésükre került be a rendszerbe egy speciális, ember által nehezen megoldható prozódiai alternatíva (a mondat szavainak kijelentő hanglejtésű, de szünetekkel elválasztott gépi felolvasása is). Noha a fejlesztések magyar nyelven folynak, kimondottan törekedtem a nyelvfüggetlen megoldásokra.

A kutatásban számos MSc és PhD hallgatóm vett részt. Közülük kiemelkedik Tóth Bálint (PhD) és Nagy Péter doktorandusz hallgató hozzájárulása.

Számszerű kiértékelés:

A beszédfelismerőt és szövegfelolvasót egyaránt tartalmazó mintarendszer tesztelése a MATÁV PKI-ban történt 2005-ben. Ennek kritikus eleme volt a beszédfelismerő, ezért az erre vonatkozó teszteredményeket ismertetem. A beszédfelismerő szótárába a cél-területhez illeszkedő 528 elemet (lehetőleg több szótagú szókapcsolatokat, pl. *Mit vegyek a boltban?*) vettünk fel. 12 személy olvasott fel ezek közül 50-50 elemet zajos (irodai) környezetben ISDN telefonon keresztül. A vonal másik végén a VoxAid2006 alkalmazás futott és visszaolvasta a felismert elemet.



32. ábra. A beszédfelismerő alrendszer tesztjének eredményei.

Ha a felismerés hibás volt, a felhasználónak meg kellett ismételni azt legfeljebb háromszor. A harmadik ismétlés után tovább kellett lépni a felolvasási sorban. A felismerés pontossága nemcsak a beszédfelismerő pontosságától függ, hanem a felismerendő szótár tervezésétől is. Tehát a teszt egyszerre vizsgálja mindkettőt.

Az eredményt a 32. ábra mutatja. Az elsőre helyesen felismert szavak aránya 95,5% volt. A második bemondásra (1. ismétlés) további 3% helyes válasz érkezett. Tehát a hibás felismerések aránya ebben a kísérletben elenyésző volt. Sajnos ez a rendszer prototípus maradt, nem került éles alkalmazásra.

A helyszűkére tekintettel csak az aktuális változatra (VoxAid2012) vonatkozóan, 13 tesztelővel végzett teszt összegzett eredményeit ismertetem a 9. táblázaton. A teszt elején mindenki kapott 20 percet, hogy megismerkedjen a rendszerrel. Az alkalmazás egyértelműségének vizsgálata érdekében a tesztalanyok csak alapvető ismertetést kaptak a rendszer működéséről (PC-s változat: VoxAidDesktop, Android telefon: VoxAidAndroid). Majd három feladatot kellett elvégezni:

a.) Megtalálni és felolvastatni egy megadott mondatot
b.) A kijelzés betűméretét egy előre megadott értékre beállítani
c.) Új kötött szövegkategóriát létrehozni és abban elhelyezni egy mondatot.

9. táblázat. Összegzett eredmények (átlag ± szórás).

	VoxAidAndroid	**VoxAidDesktop**
	Érték*	Érték*
Feladat megoldhatósága	4,71 ± 0,24	4,83 ± 0,17
Kezelhetőség	4,71 ± 0,24	4,67 ± 0,27
Logikus szerkezet	4,71 ± 0,24	4,83 ± 0,17
Futási sebesség	5 ± 0	4,83 ± 0,17
Funciók elérhetősége	4,14 ± 0,14	4,5 ± 0,3
Használhatóság	4,85 ± 0,14	4,5 ± 0,7

*1: használhatatlan, 5: tökéletes

A 9. táblázat első sora a feladatok egyszerű megoldhatóságára vonatkozott. A nagyobb eszközön értelemszerűen ez könnyebben ment. Az asztali változat több funkcióval rendelkezik, ezért nehezebb kezelni. Bizonyos funkciók (pl. kategória szerkesztése) az Android változat kisebb képernyője miatt a menüben mélyebben helyezkednek el, ezért rosszabb a szerkezeti és az elérhetőségi osztályzata. A futási sebességre nagy hangsúlyt helyeztünk, mert valós idejű kommunikációnak a késleltetés alapvető korlátja lehet. Valószínűleg más alkalmazásokhoz képest tett az okostelefonos változat jó benyomást a tesztelőkre ezen a térem. A mobilitás fontosságát kiemelték a tesztelők, ez lehet az oka az okostelefonos megoldás előnyének a használhatóság terén.

Konklúzió:

Az általam kidolgozott VoxAid/StrokeAid eljárás alkalmasságát multimodális felhasználói felületek beszédsérült emberek kommunikációjának rehabilitációjára részben prototípusok tesztelésével, részben 8 logopédus szakmai gyakorlatában és három beszédsérült ember mindennapi életében igazoltam.

9. Az eredmények alkalmazásai, műszaki alkotások

A téziseimben bemutatott új kutatási eredmények gyakorlati alkalmazásokban és műszaki alkotásokban is felhasználásra kerültek. Ebben a fejezetben négy alapvető felhasználási területet tekintek át. A közcélú beszéd-interakciós rendszerek kizárólag beszéd modalitást felhasználó megoldások, jellemzően távközlő hálózaton vagy közlekedési utastájékoztató rendszerekben. Az egészségügyi alkalmazásokról szóló alfejezetben új, innovatív, többféle modalitást kombináló megoldásokat ismertetek. A fogyatékos és idős emberek számára fejlesztett, valamint az általános célú rendszereinket a terjedelmi korlátokra tekintettel csak felsorolom.

9.1. Közcélú beszéd-interakciós rendszerek

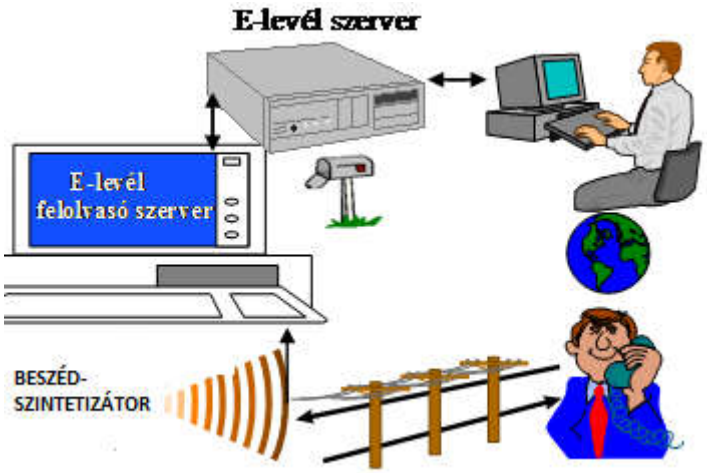
A gépi beszédkeltés kutatásának célja, hogy az eredményeket az adott kor technológiai lehetőségeinek megfelelően minél szélesebb felhasználói körhöz eljuttassa a gyakorlatban felhasználható szolgáltatásként. Ebben az alfejezetben a közcélú felhasználásra tervezett és megvalósított olyan rendszereket tekintjük át, melyek az értekezés kutatási eredményeire épültek.

9.1.1. Elektronikuslevél-felolvasó rendszer távközlési szolgáltatásként

(MailMondó, 1999-[1]), [104], [105], [49], [48])

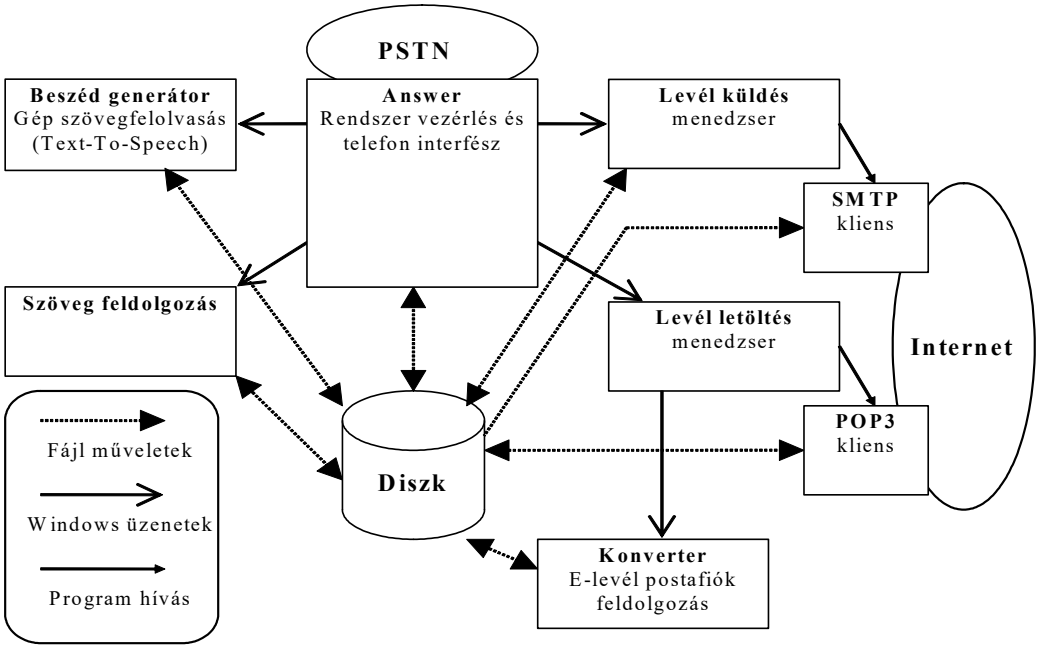
Az elektronikus levelek hozzáférhetősége kézenfekvő igény gépi felolvasással, az okostelefonok korában is, például gépkocsivezetés közben. A 90-es évek végén, amikor egy ilyen megoldás fejlesztési ötlete felmerült, a számítógépek hozzáférhetősége sokkal korlátozottabb volt, mint ma. Viszont a vezetékes és a mobiltelefonok gyakorlatilag minden nagykorú magyar állampolgár számára hozzáférhetőek voltak már akkor is. Először kutatási infrastuktúraként alakítottunk ki egy prototípust [104]. Ebből fejlesztettük tovább az I. téziscsoportban szereplő eredményekre építve a világ egyik első ilyen célú hálózati szolgáltatását, ami a legnagyobb magyar távközlési szolgáltató éves jelentésében innovációs eredményként jelent meg [49]. A megoldás általános felhasználói blokkdiagramja a 33. ábra alapján követhető. Az ábra jobb felső szélén láthatjuk az elektronikus levél feladóját. Alul pedig az üzenetet meghallgató partnert.

[1] http://kutyu.hu/cikk/3931/ (Westel 900: bemutatkozik a Mailmondó)



33. ábra. Elektronikuslevél-felolvasó általános blokkdiagramja
[105]

Az e-levél szerver fogadja a leveleket. Az e-levél felolvasó szerver alakítja át a strukturált dokumentumot felolvasható szövegállománnyá, amit (ebben az első változatban a MultiVox, később a ProfiVox) gépi szövegfelolvasó rendszer alakít át hanggá és juttat el a címzetthez. Ebben az időszakban a gépi beszédfelismerés még nem volt magyar nyelven erre a célra alkalmazható, így a választ automatikusan generált e-levél mellékleteként küldte el a rendszer. A melléklet a bemondott üzenet felvételével és „becsomagolásával" áll elő. Az egyidőben 30 felhasználót kiszolgáló rendszer egy 286Mhz-es Pentium II processzoros, 64Mbyte RAM-mal és Windpws NT 4.0 operációs rendszerrel ellátott számítógépen futott. A telefonos interfészt egy Dialogic 2Mbit-es PC kártya biztosította.



34. ábra. E-levél felolvasó szoftver architektúra
[105]

A 34. ábra mutatja a megoldásunk szoftver architektúráját. Az Answer modul végzi a rendszer elemeinek koordinálását és a telefonos interfész kezelését. A *Letöltés menedzser* vezérli a felhasználó által meghallgatni kívánt levelek hozzáférését a postafiókban (később a POP3 mellett IMAP interfész is készült). A *Konverter* modul a postafiók tartalmát elemzi, és különválasztja a felolvasáshoz szükséges elemeket (feladó, levél tárgya, a levél törzse, mellékletek, stb.). A *Szövegfeldolgozás* alrendszer felelős az elektronikus levél felolvasható formába hozásáért. A feladó címe szinte reménytelen feladatot jelent, hiszen az többnyire nem tartalmaz magyar nyelvű értelmes elemeket. A levél tárgya és törzse is számos kihívást tartogat. Első lépésként meg kell határozni a szöveg nyelvét. Az első változatban magyar, angol és német, ezt mondatonként végeztük, később a szó szintű megoldást is kidolgoztuk akár 77 nyelvre is [106]. A kutatás során alapvető nyelvstatisztikai vizsgálatokat is végeztünk [107]. Az 1990-es évek végén még gyakran írták a leveleket az angol ASCII kódkészlet betűivel. Ez a karakterkészlet nem tartalmazza a magyar ékezetes betűket.

Ez számos félreértést okozhat. Ami az írott formából általában könnyen kikövetkezhető, az a hangzó változatból nehezen kezelhető. A 10. táblázat erre ad érzékletes példát. Ezért a második lépésben vissza kell állítani az ékezetes alakokat. Gyakorisági elemzés alapján 95% feletti pontossággal tudtuk megoldani ezt a feladatot. Kritikus elemet jelentettek a személynevek. Például *Veres Péter* gyakorisági alapon *Véres Péter* vagy *Verés Péter* alakra is hozható. Ezért azt a megoldást alkalmaztuk, hogy a személyneveknél elhagytuk az ékezetesítő algoritmust, feltételezve, hogy az ismerős személyt a felhasználó amúgy is azonosítani tudja. Itt történik a rövidítések és speciális karaktersorozatok (pl. e-levél és honlap címek) feloldása is.

10. táblázat. Az ékezetek jelentésmódosító hatása az „*agyat*” karaktersorozat esetén [105]

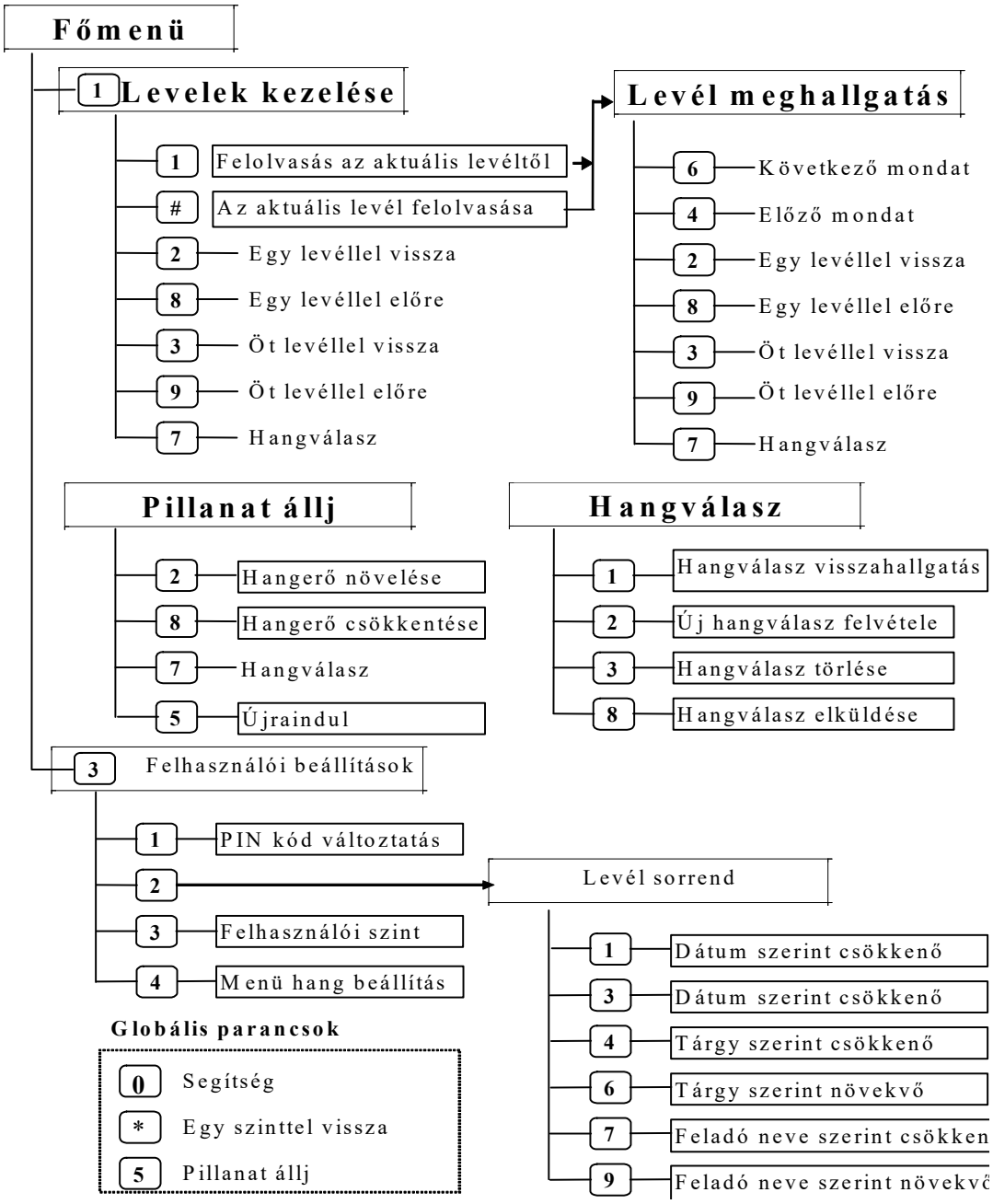
agyat	**Agyat** operálni veszélyes dolog.
ágyát	Megvetette az **ágyát**
agyát	Elborította az **agyát** a düh.
ágyat	**Ágyat** szeretett volna venni a bútorboltban.

A *Beszédgenerátor* modul a *Konverter* által előállított normalizált szövegből állítja elő a beszédet, amit az Answer alrendszer továbbít a telefonvonalon keresztül a felhasználó felé. Amennyiben előre generált szöveges, vagy hangfelvétel melléklettel ellátott választ kíván küldeni a feladónak, a *Levélküldés* egység kerül aktiválásra.

Az egyes modulok többféle módon kapcsolódhatnak egymáshoz. Mivel általában jelentős méretű adatállományok kezeléséről van szó, az adatcserét jellemzően a *Diszk* tároló rendszeren keresztül végzik a rendszer komponensek. A vezérlési műveletek pedig Windows üzeneteken

vagy programhíváson keresztül valósulnak meg. A moduláris felépítés sok előnnyel jár. Egyrészt egyszerűbb a rendszer elemeinek (akár üzem közbeni) frissítése, másrészt a 24 órás üzem miatt szükséges automatizált rendszerfelügyelet is könnyebben megoldható.

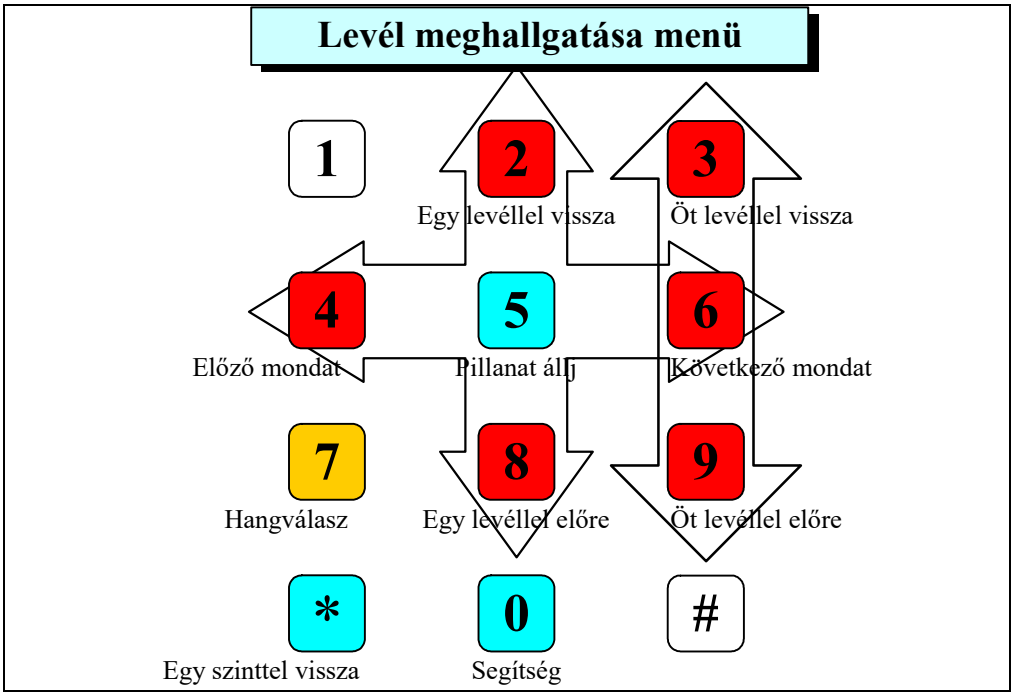
Megoldásunk meghatározó fontosságú eleme az ember-gép kapcsolat tervezése. A kor színvonalán megbízhatóan csak nyomógombos vezérlés volt alkalmazható. Az interaktív hangválasz rendszerek (IVR) sokszor átgondolatlan menürendszere a mai napig sok nehézséget okoz az ügyfeleknek [108].



35. ábra. Az e-levél felolvasó hangos menü szerkezete (az azonosítási eljárás nélkül [48]*)*

Saját megoldásunkat a 35. ábra mutatja. A menü szerkezetét a nyomógombos telefonokon jól azonosítható (kis dudorral megjelölt) 5-ös gomb köré terveztük. Ezzel lehetővé vált, hogy a nyomógombos (mobil)telefonokkal anélkül lehessen a rendszert használni, hogy ránézzünk a készülékre. Az 5-ös gomb egyben globális parancsként (bárhol megnyomva, ugyanazt a hatást éri el) *Pillanat állj* funkciót lát el. Leáll az aktuális futó funkció/bemondás. A sorban fölötte levő gombbal (2-es) a hangerőt növelni, az alatta levővel (8-as) csökkenteni lehet. A gomb újabb megnyomásával visszatérhetünk a megszakított funkcióhoz. Így például meg lehet ismételtetni egy nem jól értett mondatot is. Amennyiben éppen e-levél kezelése folyt a *Pillanat állj* funkció aktiválásakor, lehetőség van a 7-es gomb megnyomásával a *Hangválasz* menübe lépni. Itt lehetőség van a válasz felvételére, visszahallgatására, törlésére és elküldésére is egy előre beállított, formalizált szöveges üzenet mellé csomagolt mellékletként.

A levelek meghallgatása két szinten valósulhat meg. A főmenüből elérhető *Levelek kezelése* menüpont kiválasztásával egymás után meghallgathatjuk a legutoljára érkezett levél feladójának nevét (vagy ha az nincs, akkor e-levél címét), a levél tárgyát és a levél küldési időpontját. Ha a teljes levelet meg szeretnénk hallgatni, akkor két lehetőség áll előttünk. A # gomb megnyomásával az aktuális levél kerül felolvasásra és utána visszakerülünk a főmenübe. Az 1-es gomb megnyomásával pedig a *Felhasználói beállítások* szerinti sorrendben meghallgathatjuk az aktuális levelet és az utána következőket folyamatosan, bármilyen másik gomb megnyomása nélkül.



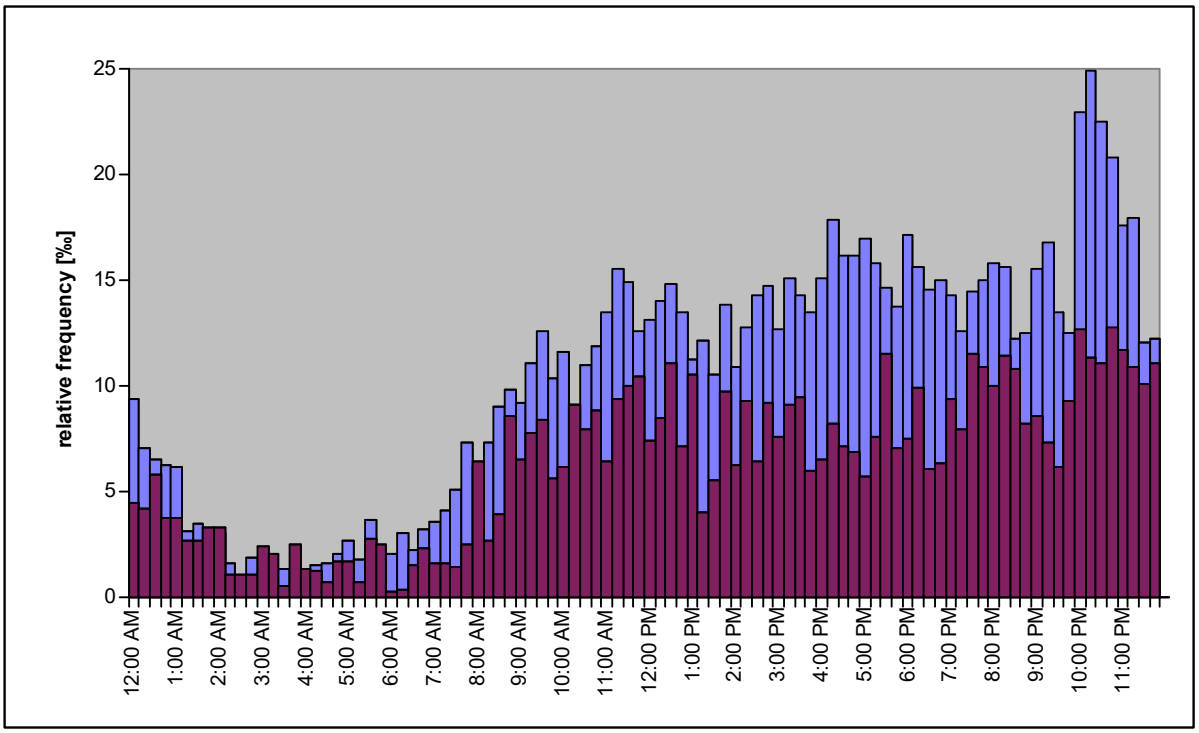
36. ábra. Függőleges (levelek között) és vízszintes navigáció (egy levélen belül) a „Levél meghallgatása” menüben [48]

A leveleken belül és a levelek között a 36. ábra szerint lépkedhetünk. Az 5-ös gombhoz képest függőlegesen elhelyezkedő gombokkal a levelek között, a vízszintesen elhelyezkedőkkel pedig az aktuális levélen belül, a mondatok között tudunk navigálni.

A Felhasználói beállítások menüben 0-6 számjegyű azonosító PIN kódot határozhatunk meg. Alapértelmezésben nincs PIN kód (0 számjegy). Szintén itt állíthatjuk be a levelek felsorolásának sorrendjét. Innovatív, ismereteink szerint korábban nem alkalmazott megoldásunk az, hogy állítható a felhasználói szint (kezdő, haladó, profi). Az automatikus hangválasz rendszerek egyik kritikus pontja az, ha túl sokat, túl részletesen magyaráz a rendszer, vagy ha túl keveset, így könnyen elveszik a felhasználó a menürengetegben. A kezdő felhasználó számára minden szinten részletes információt ad a rendszer. Minden gombnyomást „Köszönöm.” üzenettel nyugtáz a rendszer. A haladó szinten ez a nyugta elmarad és rövidebbek a rendszerbemondások. A profi szintű felhasználókat csak rövid, néhány szótagos emlékeztetőkkel segítjük. A Menü hangbeállítási lehetősége szintén új, innovatív megoldásunk. A felhasználó kiválaszthatja, hogy férfi vagy női hangon szeretné a rendszerüzeneteket meghallgatni. Általános vélekedés volt, hogy a felhasználók megszokták, hogy egy ilyen rendszerben női hang szól. A gyakorlatban kiderült, hogy jelentős számban átállították a TTS hangjához hasonló férfihangra.

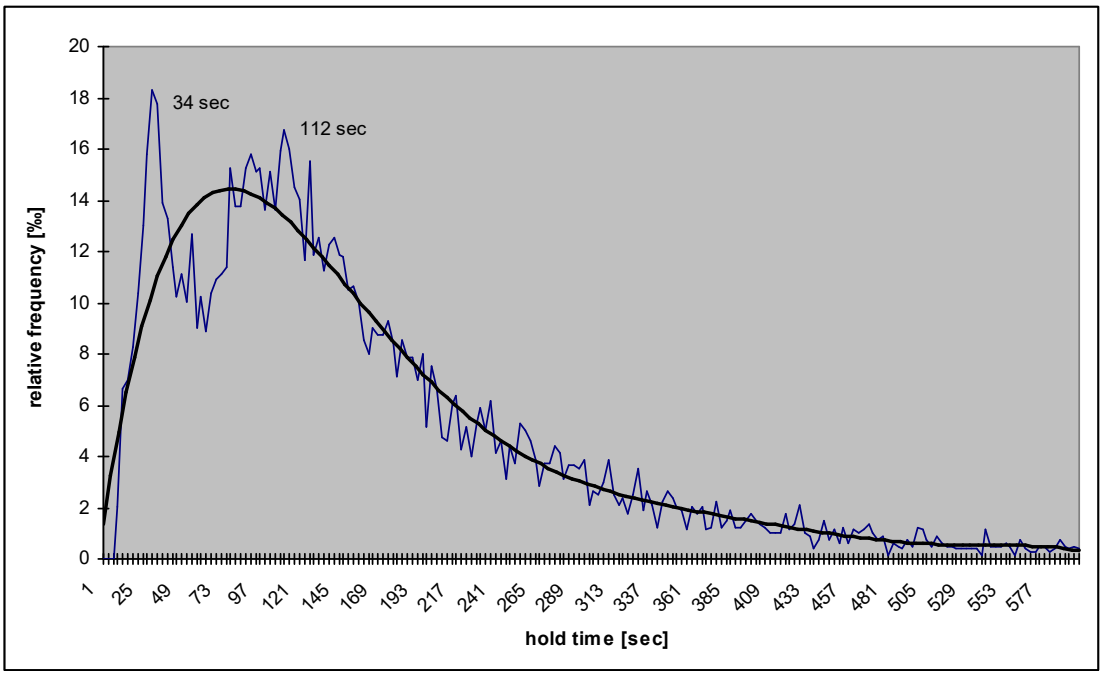
Érdekes tapasztalatokat gyűjtöttünk a valós felhasználók *Felhasználói beállítások*kal kapcsolatos aktivitásának vizsgálatával. Az “éles” működés első hét hete alatt legaktívabb (a legtöbb időt a rendszerben töltő) 300 felhasználó tevékenységét elemeztük. Mintegy háromnegyedük legalább egy opciót megváltoztatott. Körülbelül 70% legalább egyszer PIN kódot változtatott. A legtöbbet változtató felhasználó hét hét alatt 28 alkalommal cserélt. Az ügyfelek nagyjából 60%-a legalább egyszer változtatta a felolvasási sorrendet és 20% legalább négyszer állított ezen. A legnagyobb változtatási adat 21, ami 10, átlagosan több, mint négy perces hozzáférés alatt ment végbe. Érdekes, hogy a felhasználók 47% megmaradt a kezdő felhasználó szinten, talán, mert igénylik a részletes tájékoztatást. Az ügyfelek 40%-a 1-4 alkalommal változtatta a felhasználói szintet. Mintegy 60%-uk állította át az alapértelmezett női rendszerüzenet hangot férfihangra legalább egyszer. 30% változtatta a rendszerhangot legalább négyszer. A hét hetes időszak után 17% használta a férfihangot. Egyértelműen látszik, hogy minden felhasználói opciót igénybe vettek az ügyfelek.

A 37. ábra mutatja az e-levél felolvasó felhasználóinak az aktivitását az „éles” üzem első 10 hete alatt. A világos oszlopok a munkanapokat, a sötétek pedig a hétvégéket jelölik.



37. ábra. A felhasználói forgalom eloszlása munkanapokon (világos) és hétvégén (sötét)
[48]

A 38. ábra alapján tekinthetjük át az „éles” működés első 10 hetében sikeresen azonosított felhasználók átlagos tartási idejét (amíg használták a rendszert). Az ábra alapján az ügyfelek két



38. ábra. A sikeresen azonosított felhasználók tartási idejének eloszlása (a simított görbe a valós adat 6-odrendű polinommal történő közelítése)
[48]

csoportba sorolhatók. Az első csoport csak arra kíváncsi, hogy érkezett-e új üzenet. Ez okozza az első csúcsot 34 másodpercnél. Jellemzően egy percen belül zárják a hívást. A második csoport már meghallgat legalább egy levelet. Ehhez tartozhat a két perc körül található csúcs. Az átlagos tartási idő majdnem három perc. Érdemes megjegyezni, hogy voltak több mint 10 percnyi tartási időt elérő felhasználók is.

Az okostelefonok korában első látásra elavultnak tűnhet ez a szolgáltatás, de gondoljunk arra, hogy számos ember számára az autó a harmadik élettér és vezetés közben a legtöbb országban tilos huzamosan a képernyőt nézni. Ezekben a helyzetekben ma is célszerű szolgáltatás az elektronikus levelek felolvasása. A mai technológiával ez nem igényel jelentős központi erőforrásokat, hanem magán a telefonon is megoldható, akár egyszerű beszédfelismerővel történő vezérléssel is.

9.1.2 SMS-felolvasó rendszer okostelefonon

(SMSmondó, 2003-, [54], [55])

Az okostelefonok használata jelentős mértékben eltér a vezetékes készülékekétől. A legfontosabb különbség, hogy az okostelefon nem helyhez kötött, hanem személyes jellegű tárgy, az emberek mindenhová magukkal viszik. A személyes jellegük azzal is jár, hogy fontos a testre szabhatóságuk, a felhasználóik életmódjához való illeszthetőségük. Ennek következtében bonyolult, összetett felhasználói felületek alakultak ki. Ezek kezeléséhez két alapvető megközelítés létezik. Egyrészt az érintőképernyő és a rajta elhelyezett szoftvervezérelt elemek (billentyűzet, ikonok, funkcióbillentyűk – soft keys), másrészt pedig a beszédtechnológiák növekvő szerepe.

A beszédtechnológiai területen a hangsúly elsősorban a gépi beszédfelismerésen van, különösen az autóvezetés biztonsági szempontjai miatt. A világ számos országában tilos a telefont kézben tartva vezetni, ezért speciális befogó szerkezeteket fejlesztettek ki, melyek lehetővé teszik a kezeket szabadon hagyó (hands-free) kezelést. Ennek ellenére a névjegy lista böngészése, a híváskezdeményezés és befejezés vagy az SMS (és a szaporodó azonnali üzenetküldő rendszerek) üzeneteinek írása és olvasása hosszabb-rövidebb ideig elvonja a vezető figyelmét az úttestről.

A fenti funkciókat jól lehet irányítani beszédfelismerés segítségével, de ezt valós autós körülmények között korlátozzák a változó akusztikai feltételek. Egy másik motivációs tényező az eszközök méretének csökkentése (pl. okosórák).

Sokkal kevesebb figyelmet kapnak az automatikus beszédkeltés lehetőségei az okostelefonokon. Ebben a szakaszban erre mutatok rá az SMS-felolvasás témaköre kapcsán.

Már hosszabb ideje rendelkezésre állnak gépi szövegfelolvasó rendszerek számos nyelven, azonban ezek jellemzően (és különösen a 2000-es évek elején) központi, nagy kapacitású szervereken futottak. A felolvasó szolgáltatás pedig kapcsolt távközlési hálózaton keresztül volt elérhető. Ez a kapcsolat lényegesen drágább, mint a csomagkapcsolt megoldás. Ezért kézenfekvőnek tűnt az a gondolat, hogy a felolvasást próbáljuk a helyben levő okostelefonon megoldani és csak a felolvasandó szöveget töltsük le valamilyen adatkapcsolaton keresztül.

A gépi beszédkeltés lehetővé teszi részletesebb kontextus-függő súgó rendszerek kialakítását is. A telefonok viszonylag kicsi képernyőjén nehézkesen fér el egy alkalmazás és a funkciót támogató súgó felülete egyszerre. A gépi szövegfelolvasás (TTS) lehetővé teszi ún. hangprofilok kialakítását (jellemzően hangmagasság, hangerő, beszédsebesség vagy akár beszédhang, beszédstílus, stb. változtatásával). A különböző üzenet típusokhoz, hívásjellemzőkhöz más-más profilt rendelhetünk, így már az üzenet első hangjainak elhangzásakor érdemi információhoz juthatunk.

Példaként tekintsünk egy olyan hölgyet, aki kisebb látásélesség probléma miatt olvasáshoz szemüveget használ. Általában a szemüveget és a mobiltelefont is a táskájában tartja.

Ha SMS üzenete érkezik,

1. ki kell, hogy vegye a szemüveget a táskából,
2. feltenni a szemüveget,
3. kivenni a mobiltelefont a táskából,
4. felnyitni a képernyőzárat,
5. elolvasni az üzenetet,
6. visszazárni a mobilt,
7. betenni a táskába,
8. levenni a szemüveget,
9. betenni a táskába.

Ha a telefonban lenne egy SMS felolvasó alkalmazás, legalább az 1., 2., 8., 9. lépések elhagyhatók lennének. Ha egyedül lenne egy csendes helyen, ahol a mobil hangja érhető lenne a táskán keresztül is, akkor a többi lépés is kihagyható, hiszen a rendszer automatikusan fel tudja olvasni az üzenetet. Hasonló logika alapján olyan szerteágazó területek is bevonhatók, mint az otthon automatizálása (riasztók, mosógép, hűtő, stb. jellemzői), autós információs rendszerek, pénzkiadó automaták, ill. bármely olyan helyzetben mikor a szemünk és/vagy a kezünk foglalt (pl. főzés, szerelés, koszos műveletek).

Végül, de nem utolsósorban a súlyosan látássérült és a vak emberek számára az alapvető hívásfogadáson túli funkciók is megnyílnak a TTS-en alapuló alkalmazások révén. A 2000-es évek elején egyetlen képernyőolvasó alkalmazás volt elérhető Symbian operációs rendszerre angolul és néhány nagyobb európai nyelvre.

A mobilkészülékek egy igen széles körben felhasználható tulajdonsága az SMS küldés/fogadás. Ennek a népszerű szolgáltatásnak azonban számos felhasználási korlátja is van, mint pl. az, amikor az SMS munkavégzés, vezetés közben érkezik, a címzett nem tudja azt rögtön megnézni, esetleg szemüveg nélkül nehezebben tudja elolvasni. A telefonhívás kezdeményezése, fogadása lehetséges telefonszám beütése nélkül (egygombos tárcsázás, hangtárcsázás, automata hívásfogadás), ám a beérkezett üzenetet a hagyományos megoldással a 2000-es évek elején még minden esetben meg kellett nyitni és szóról-szóra végigolvasni. Ez kifejezetten veszélyes és tilos gépjárművezetés közben, hiszen több másodpercre elveszítjük a kapcsolatot a környezetünkkel, sőt a közeli tárgyra való fókuszálás további értékes pillanatokat pazarol el egy veszélyhelyzet felismeréséből.

A szövegek beszédhanggá alakítása nagy számítástechnikai kapacitásokat is igényelhet, és ezért az SMSmondó rendszer fejlesztésekor (2003) nem volt nyilvánvaló, hogy az erre alkalmas szoftver mobilkészülékbe tölthető legyen. Az emberi hangból tárolt hangmintákra épülő eljárások igen nagy memóriaigényűek lehetnek, míg a teljesen szintetizált hangok az emberi hangtól igen távol esnek, robotosak. A szoftver mérete, CPU igénye és a hangminőség között tehát optimális fejlesztési kompromisszumot kell kötni, új megoldásokat kell kidolgozni, hogy végül a szöveg jó minőségű feldolgozása a mobiltelefonba beépíthető legyen. Az I. téziscsoportban ismertetett kutatási eredményekre építve dolgoztuk ki mobiltelefonra optimalizált megoldásunkat. Ma már hasonló alkalmazásunk a III. téziscsoport szerinti eredmények alapján is elérhető.

Az SMSmondó alkalmazással a fenti korlátokat oldottuk meg (a M.I.T. Systems Kft-vel együttműködésben), tudomásunk szerint a világon először. A legnagyobb magyar távközlési szolgáltató éves jelentésében innovációs eredményként jelent meg [54], 2004-ben pedig az Innovációs Nagydíj pályázat informatikai területének I. helyezettje lett. Az igényt mi sem jelzi jobban, mint hogy a termék megjelenése után már néhány nappal Internetes fórumok tárgyalták a szoftver feltörésének lehetőségét.

39. ábra. Az SMSmondó alkalmazás felhasználói környezete

Az alkalmazás lényegét a 39. ábra mutatja be (az egyszerűség kedvéért a telefont tartó konzolt elhagytuk). Az SMSmondó a világon az első olyan telefonkészüléken futó alkalmazás volt, amely a felhasználó beállításainak megfelelően képes a beérkező üzenetet felolvasni. Felhasználóbarát mivoltát az is jelzi, hogy kezeli a rövidítéseket, csupán egyszer meg kell adnunk a megfeleltetéseket.

A Symbian okostelefonok rendelkeztek beépített kihangosítóval, az alkalmazás nem csak autóvezetés közben hasznos, hanem minden olyan helyzetben, amikor valami miatt képtelenek vagyunk az olvasásra (gyenge/erős fényviszonyok, fontosabb tevékenység), esetleg a gombok nyomkodására (szennyezett ujjak). Például egy sietős gyaloglás során nagyon kényelmes, ha a telefonunk az ingzsebből adja tudtunkra az üzenet tartalmát, miközben mi figyelhetünk a többi gyalogostársunkra, vagy az úttesten való biztonságos átkelésre.

Az operációs rendszer szerinti *Általános, Csend, Megbeszélés, Kültéri* és *Pager* felhasználói profilokhoz az SMSmondóban is egyéni beállításokat lehet meghatározni. Egy gombnyomással lehet az *Üzenetek felolvasása* automatikus. Az *Érkező üzenet jelzése* opció bekapcsolásakor üzenet érkezésekor az "*Önnek új üzenete érkezett*" előre definiált bemondás hangzik fel. A *Feladó* neve (ha a névjegyzékben elérhető) vagy telefonszáma *felolvasása* is beállítható valamint a *Feladási időpon*t is. Az SMSmondó ablak *Automatikus kikapcsolás*a is megoldható 5 vagy 10 másodperc inaktivitás után. Az üzenetet fel lehet olvastatni csak egyetlen egyszer, vagy beállítható

az ezután következő *Ismétlések száma*. A beszéd *Hangerő, Beszédsebesség* és *Hangmagasság* jellemzői is konfigurálhatóak minden egyes felhasználói profilban külön-külön.

Érdemes a hangerőt alacsony értékre állítani vagy teljesen lenullázni a *Csend* és a *Megbeszélés* profiloknál. A *Kültéri* profil esetében a nagyobb hangerő, az alacsonyabb beszédsebesség és a dupla ismétlés segíthet az üzenet megértésében zajos környezetben. A *Pager* profil mellett érdemes az *Üzenetek felolvasása* funkciót kikapcsolni. Ha például gyereket szállítunk az iskolába és nem akarjuk, hogy az üzenetünket hallja, akkor nem nyomunk meg egy gombot sem az üzenet érkezésekor. Miután kitettük a gyereket az iskolánál egy gombnyomással meghallgathatjuk. A rendszer telepítésekor a profilok hozzájuk illeszkedő alapbeállításokat kapnak.

A rövidítések és a feloldásuk (ill. bármilyen karaktersorozat és a hozzájuk tartozó kiejtés) szabadon megadhatók. A kiejtést úgy definiálhatjuk, hogy olyan szöveget írunk be a felolvasáshoz, amit a hozzá rendelt karaktersorozat mellett hallani akarunk. A korábban bevitt rövidítési lista elemeit módosíthatjuk vagy törölhetjük is.

Az alkalmazást regisztrálni kell felhasználás előtt a mobilszolgáltatónál. A felhasználónak mindössze a nevét kell megadnia és az ehhez generált kódot kell elküldeni SMS-ben a szolgáltatónak. A válaszul kapott regisztrációs kódot pedig be kell írni az alkalmazásba és az máris használatba vehető.

A tervezés során alapvető cél volt, hogy egyrészt a felhasználók minél jobban testre szabhassák a rendszert. Másrészt a menüelemek számát igyekeztünk a lehető legalacsonyabban tartani, hogy ne bosszantsuk az ügyfeleket a túl sok beállítással. A felhasználó-orientált tervezési módszertan keretében felhasználói tesztek alapján határoztuk meg az alapbeállításokat

A rendszer sikeresen működött a 2000-es évek végéig, amikor a Symbian operációs rendszer elvesztette piaci részesedését. Az Android és az Apple operációs rendszerekre a későbbiekben készültek hasonló alkalmazások, de a fent ismertetett innovatív funkciókkal tudomásunk szerint egyik sem rendelkezik.

9.1.3 Egy távközlési szolgáltató árlistabemondó szolgáltatása

[58]

A távközlési szolgáltatók általában emberektől előre felvett rendszerüzeneteket (ún. prompt-okat) használnak telefonos információs rendszereikben. Több száz ilyen üzenet felvétele még egy tapasztalt rádióbemondó számára is meglehetősen fárasztó és időt rabló. A mindenkori bemondás, mint megoldás egyben meglehetősen költséges és sok egyéb tényezőtől függ (a bemondó személy és a hangstúdió elérhetősége, a bemondó személy egészségi állapota, stb.). Kézenfekvő alternatíva

az emberi hangfelvétel (statikus elemek) és a gépi beszédkeltés kombinálása (TTS a változó elemekre). Viszont ekkor biztosítani kell a hangzás folyamatos jó minőségét. Korábbi megoldások [109], [110] ezt úgy valósították meg, hogy egy beszédtechnológiai szakértőnek kellett optimalizálást végezni a beszédelemek kiválasztása során

A II. téziscsoportban ismertetett kutatási eredményeinkre alapozva és egy mobil távközlési szolgáltatóval együttműködve egy lényegesen hatékonyabb módszert dolgoztunk ki. A korlátozás abban áll, hogy a tématerület szűk: a szolgáltató ajánlott készülékeinek hetente változó árlistáját kell automatikusan generálni a megadott írásos dokumentum alapján. Kiinduláskor rendelkezésre állt a cég által foglalkoztatott női bemondó korábbi felvételeinek szöveges listája és a témakörhöz kapcsolódó hangfelvételek. A mintegy három év alatt 3747 mondat került felvételre, ami 69.057 szót tartalmazott, és összesen 8 óra és 25 perc hosszú hanganyagot tett ki. Néhány jellemző mondatot és a hozzájuk tartozó szöveges átírást láthatunk az alábbiakban:

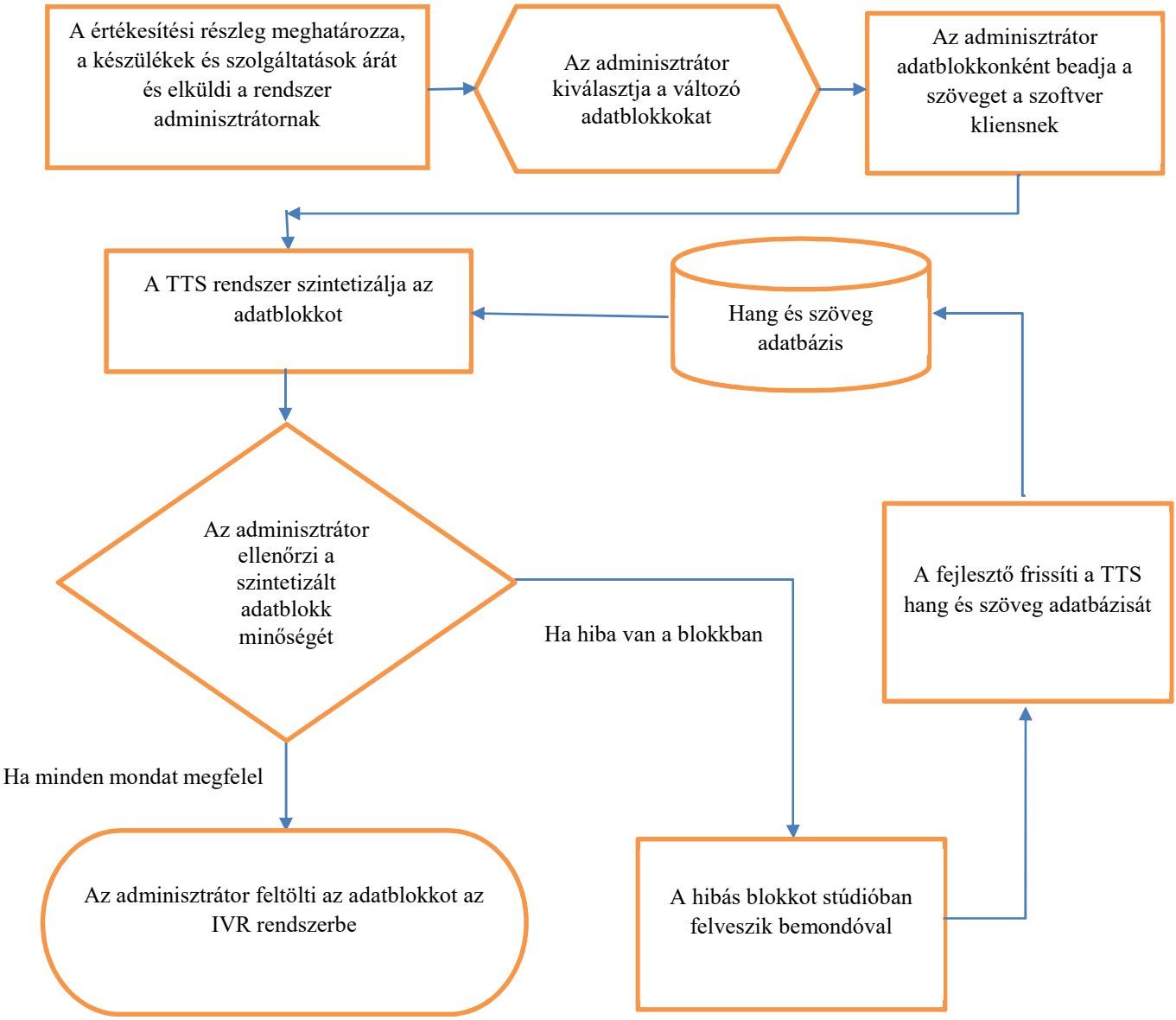
LG Shine KE970 = “el dzsí sájn ká ee kilencszázhetven”
Web'n'walk Box 7.2+ = ”veb en vók boksz hét pont kettő plusz”
SonyEricsson C902 James Bond Edition = ”szoni erikszon cé, kilencszázkettő dzsémsz bond edisön”

A fejlesztés során meg kellett oldanunk a korábbi felvételek szövegének és hangjának hangszintű szinkronizálását és egységes adatbázisba rendezését és a fentebb látható szövegátalakítási feladatot. Ehhez az I. tézis szerinti ProfiVox rendszer graféma-hangkód átalakító moduljából indultunk ki és azt egészítettük ki ehhez a feladathoz tartozó szabályokkal. A cég üzleti folyamataiba illesztett rendszer működése a 40. ábra alapján követhető.

A cég értékesítési részlege (jellemzően a hét végén, pl. pénteken 14 órakor) meghatározza, hogy a következő héten milyen termékek és szolgáltatások milyen áron és feltételek mellett lesznek elérhetők. Ilyenkor megjelenhetnek új termék- vagy szolgáltatásnevek is. Az interaktív hangválasz (IVR, Interactive Voice Response) rendszerben hétfőn 0 órára rendelkezésre kell, hogy álljon minden új információ hangban. Az IVR rendszer-adminisztrátor megkapja a listát és annak alapján eldönti, hogy melyek az új, korábban még nem használt szövegblokkok. Ezeket a szövegblokkokat (egy vagy több mondat) beadja a II.2 tézisben ismertetett TTS rendszernek.

A rendszer által generált hangminta minőségét is az adminisztrátor ellenőrzi. Ezt a költségfüggvényből származtatott numerikus adat is segíti. Ez egy kritikus pont, hiszen elvárás, hogy az átlagos felhasználónak ne tűnjön fel, hogy vannak teljesen ember által bemondott és gép által generált üzenetrészek is a felolvasott mondatban. Ha minden rendben van, akkor az eredmény feltölthető az éles IVR rendszerbe. Ha van hibás blokk, akkor azt a cég által biztosított

stúdióban felveszik a bemondóval. Ezzel az új felvétellel kiegészítik a rendszer szöveges és hang adatbázisát.

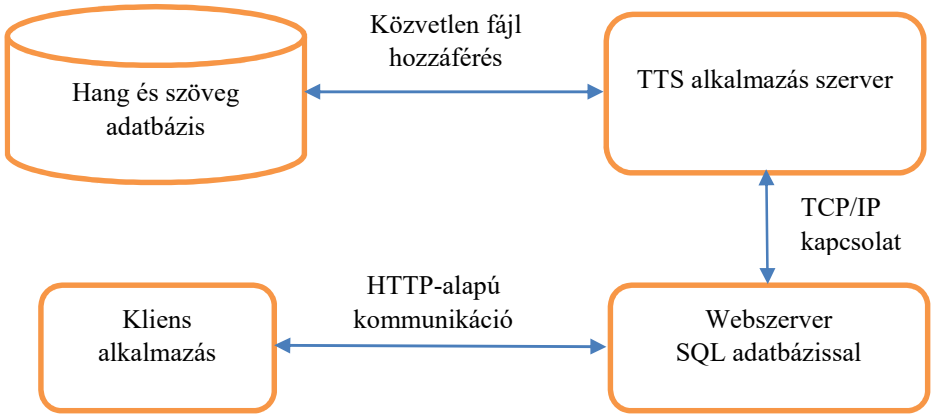


40. ábra. Az árlista gépi felolvasásának folyamatábrája

A rendszer egy webes felületen keresztül bárhonnan elérhető. A hálózati architektúrát a 41. ábra mutatja be. A TTS alkalmazás szerver és a Webszerver célszerűen fizikailag közel helyezkedik el, viszont a kliens alkalmazás bárhonnan futtatható.

Néhány heti felhasználás után az adminisztrátor a szöveg alapján nagy biztonsággal meg tudta mondani, hogy szükséges-e új hangfelvétel. Megoldásunkkal a korábban rendszeresen 2-3

személy által végzett 3-4 órás munka jellemzően fél órára egyszerűsödött. Az új blokkok mintegy 80%-át a rendszer generálta[2].



41. ábra. Az árlista felolvasás kommunikációs sémája

9.1.4. MÁV állomások hangos utastájékoztató rendszere

[111], [33]

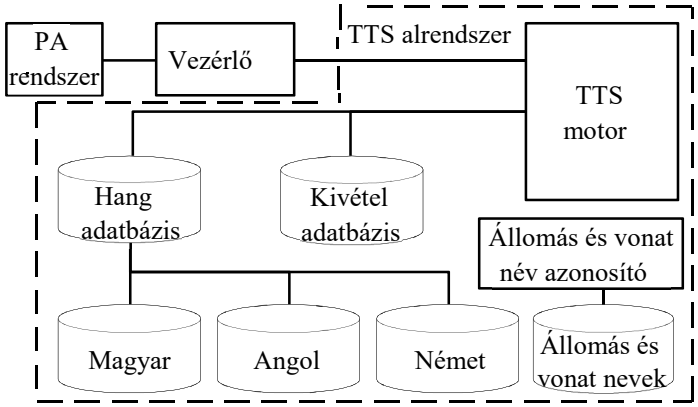
A pályaudvari hangos utastájékoztatás hozzátartozik a vasúti közlekedés minőségi utaskiszolgálási követelményrendszeréhez. Mindig volt ilyen szolgáltatás a kornak megfelelő technikai szinten, pl. [112]. A jelen szakaszban ismertetett megoldást a beszédtechnológia fejlődése, egyben a II. téziscsoportban ismertetett kutatási eredmények tették lehetővé. Rugalmasabb kezelést biztosít, olcsóbb az üzemeltetése, ugyanakkor stabilabban szolgáltat jól érthető, szívesen hallgatott hangminőséget.

A MÁV-val egyeztetve a következő követelményeket fogalmaztuk meg az új rendszerrel szemben:

- Jobb vagy legalább azonos TTS hangminőség témakör-specifikus üzenetek esetén, mint a hagyományos kézi összefűzéses rendszerben. Az idegen nyelveken elfogadható a magyar akcentus.
- Szövegbevitelen alapuló gyors (lehetőleg valós idejű) bemondás generálás.
- Érthető beszédminőség még a témakörön kívül eső üzenetek esetén is.
- Egyetlen bemondó minden nyelven.
- Az alapértelmezett magyar nyelv mellett angol (később német is) bemondások az Intercity és a nemzetközi vonatokhoz.

[2] Hangminta: http://smartlab.tmit.bme.hu/alkalmazas-arlista-felolvasas

- Helyben elérhető beszédtechnológiai és nyelvi rendszertámogatás.



42. ábra. A hangos pályaudvari utastájékoztatás rendszerterve nagy állomásokon [33]

A rendszer felépítését a 42. ábra alapján mutatjuk be. Ezt a konfigurációt nagy állomásokra terveztük. Az első ilyen konfiguráció 2014 júniusában a Budapest Keleti-pályaudvaron állt üzembe.

A hangos utastájékoztató rendszer három nagy alrendszerből épül fel: (i) a *Vezérlő* modulon keresztül tud a kezelő utasításokat adni, hogy a vasúti menetrendet és a vasúti tiszt által összeállított bemondásütemezést figyelembe véve éppen milyen hangüzenetet mondjon be a rendszer, (ii) a közcélú hangrendszer (*PA rendszer* az ábrán) tartalmazza a hangátvitelhez szükséges kábelezést, erősítőket és hangszórókat/hangoszlopokat, (iii) a szaggatott vonallal körülvett *TTS alrendszer* pedig Vezérlőtől kapott szöveget alakítja beszéddé.

Szövegkorpusz tervezés

Az elemkiválasztásos, korpusz-alapú rendszerek kritikus eleme a bemondó által felolvasandó szövegkorpusz. Egyrészt az adott témakört, másrészt az adott nyelv hangjait is a lehető legjobban célszerű reprezentálni. A kiindulási szövegállományunk egy 2007-es kísérleti rendszerhez készített listából és néhány nagy vasútállomás hagyományos bemondásainak szövegéből mohó algoritmussal származtatott tömörített anyagból állt. Az első magyar változat 2410 mondatból állt. Az ország nagyobb részének lefedéséhez 900 új mondattal kellett bővíteni. Az angol változat 577 témakör-specifikus és további általános fedést biztosító 1133 mondatot tartalmaz az ARCTIC adatbázis [113] szerint. Az üzenetek megfogalmazásánál figyelemmel kell lenni arra, hogy a magyarországi állomásokon megforduló külföldi utasok jelentős része korlátozott angol nyelvtudással rendelkezik (pl. "*The train calling at Szob...*" helyett "*The train stopping at Szob…*" a javasolt bemondás).

Különösen fontos az állomásnevek helyes kiejtése. A magyar bemondásokban az adott külföldi ország hivatalos nyelvének megfelelő kiejtést kell alkalmazni (pl. *Villach*), kivéve, ha az adott helységnek van történelmi magyar neve (pl. a szlovák *Bratislava* helyett *Pozsony*). Az angol és a német bemondásokban a magyar állomások nevét magyarul mondjuk, minden másikat az adott ország hivatalos nyelvén (pl. *The train arrives from Warszava*). A rendszerbe 2031 magyar és 732 külföldi állomásnevet vettünk fel.

A rendszerbe kezdetben 143 vonatnevet illesztettünk. Ezeket is külön jelöljük mind az írott, mind a hangzó formában. A mondat elején (*PTE Intercity train arrives from Pécs at platform 10.*) és a közepén is előfordulhatnak (*We inform our passengers that the PTE Intercity train is delayed.*). A marketing megfontolásoknak megfelelően ezek a nevek gyakran változnak. Például egy év alatt 11 új név jelent meg és 7 megszűnt.

Hangfelvételek és hangadatbázis

A hangfelvételek elkészítéséhez az első lépés a megfelelő bemondó kiválasztása. Női bemondót kerestünk, hogy így is csökkentsük a visszhangot a nagy pályaudvari csarnokokban. Három, nehezen összeegyeztethető szempontot kellett figyelembe venni, a beszédtechnológiai feldolgozáshoz és algoritmusokhoz illeszkedést, a szubjektív benyomásokat és a menedzsment szempontokat. Elég fiatalnak kell lennie ahhoz, hogy előreláthatólag a következő években se változzon a hangja. Fontos, hogy legyen tapasztalata a stúdiófelvételekkel és rendelkezzen minél sokoldalúbb nyelvtudással. Több körös tesztelés után Mátyus Katit, a Kossuth Rádió bemondó-szerkesztőjét választottuk ki, aki anyanyelvi szinten beszél magyarul és románul és akcentussal, de elfogadhatóan olvasta fel az angol és a német szöveget és a szlovák, lengyel, cseh, orosz, stb. állomásneveket is.

Annak érdekében, hogy a különböző időpontokban készített felvételek hangszíne, hangmagassága és stílusa azonos legyen, egy ún. mestermondatot alkalmazunk. Ez egy meglehetősen hosszú (11 szavas) mondat és állomásneveket is tartalmaz. Ezt a mondatot minden 25. felolvasott mondat után bejátszottuk a bemondónak fejhallgatón és meg kellett ismételnie azonos hangmagassági, hangszínezeti és beszédtempó jellemzőkkel, majd ezzel a hanggal folytatta a felolvasást. Ennek a módszernek köszönhető, hogy még ma (2019) is ugyanolyan hangszínezetű, beszédstílusú pótlólagosan kért hangfelvételeket illesztünk be a rendszerbe, mint amilyenek 2014-ben készültek.

A hangfelvételeket a szöveggel fél-automatikus módszerrel hoztuk szinkronba. Először egy ún. kényszerített felismerés üzemmódban (tudta, hogy mit kell felismerni, a megfelelő szó és hanghatárokat kellett meghatározni) működő beszédfelismerő [114] futott le. Az eredményt a hangidőtartam alapján kijelölt kézi korrekciókkal pontosítottuk. A magyar kezdeti adatbázis kb.

8 órányi beszédet tartalmazott, ami mára (2019 augusztus) mintegy 10 órára nőtt Az angol változat kb. 2 óra hosszú.

A pályaudvari utastájékoztató TTS rendszert az IT.DOT Kft.-vel együttműködve helyeztük üzembe és biztosítunk hozzá rendszertámogatást. Az első helyszín Magyarország legnagyobb személyforgalmú pályaudvara, a Keleti pályaudvar volt, 2014 júniusában. Külön kihívást jelentett, hogy a vizuális kijelzők cseréje alatt csak a hangos utastájékoztatás működött, ezért annak érthetősége kiemelten fontos volt. A rendszer sikerét mutatja, hogy ma már a Nyugati pályaudvar kivételével minden nagy forgalmú budapesti állomáson és számos vidéki csomópontban is a BME TMIT TTS megoldása működik. Azóta is folyamatos a Keleti pu. munkatársaival az egyeztetés és a kisebb-nagyobb fejlesztések elvégzése.

9.2. Egészségügyi alkalmazások

A beszédtechnológiának és ezen belül a gépi beszédkeltésnek számos alkalmazási lehetősége van az egészségügyben. Az orvosbiológiai mérnökképzés keretében több szakdolgozatot is konzultáltam ebben a témában (pl. az intenzív terápiában megnyíló témákról). A jelen szakaszban azt a két területet mutatom be, ahol nemzetközileg is újszerű alkalmazásokat hoztunk létre.

9.2.1. Magyarul beszélő NAO robot alkalmazása kórházi környezetben

[115]

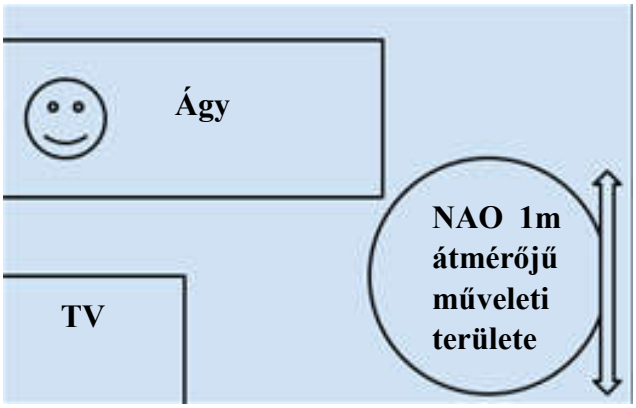
A kognitív infokommunikáció területén végzett kutatásunk [116] kapcsán került laborunkba egy NAO robot [117]. Gondolkoztam, hogy mire lehet értelmesen használni egy ilyen érdekes, de költséges (2011-ben kb. 10.000 €) eszközt. Az egészségügy merült fel olyan területként, ahol nagy hozzáadott értéket jelenthet. Némi kapcsolatkeresés után találtunk rá a Szt. László Kórház (ma Dél-Pesti Centrumkórház) Gyermekhematológiai és Őssejt-transzplantációs Osztályt[3] vezető Kriván Gergely főorvos úrra. Ő és kollégái nagyon nyitottan álltak az együttműködéshez. Így azt a célt tűztük ki, hogy megvizsgáljuk, lehet-e és ha igen, hogyan hasznosítani egy robotot a gyermek csontvelő transzplantációs eljárás folyamán. Ilyen kutatásról a mai napig sincs tudomásunk.

A célunk az, hogy a beteg gyermekek kezelését elősegítő megoldást fejlesszünk ki. A csontvelő transzplantációs szcenárió több kutatási szempontból is előnyös. Egyrészt a robotot viszonylag kis magassága (57cm) miatt a kisebb gyermekek is félelem nélkül fogadják [118]. Másrészt a

[3] https://www.gyermekdaganat.hu/mgygyt/szent-laszlo-korhaz/

kezelés elég hosszú (80-120 nap), így hosszútávú interakció is értékelhető. Kognitív kísérletek is folytathatók azzal kapcsolatban, hogy a robot inkább az intézmény képviselője, vagy inkább a gyerek társa szerepét tölti be. Hosszabb távon akár távoktatási kísérletekbe is bevonható, mert az intézmény pszichológusa szerint az egyik legerősebb motiváció a gyerekek számára a korábbi osztályközösségükbe való visszakerülés, az évvesztés elkerülése.

A NAO műanyag borítása jól tisztítható, ami a kezelés követelményei miatt fontos. A betegek 4-6 m^2-es steril helyiségekben töltik napjaikat, mivel ez a kezelés komoly kockázatokat hordoz. Ez jelentős korlátokat jelent. Egyrészt NAO csak a 43. ábra szerinti korlátozott területen belül mozoghat. Viszont vezérlése lehetőleg a helyiségen kívülről történjen, hogy a fertőzési kockázatokat így is csökkentsük. Jobb, ha a gyerekek sem érintik meg a robotot.



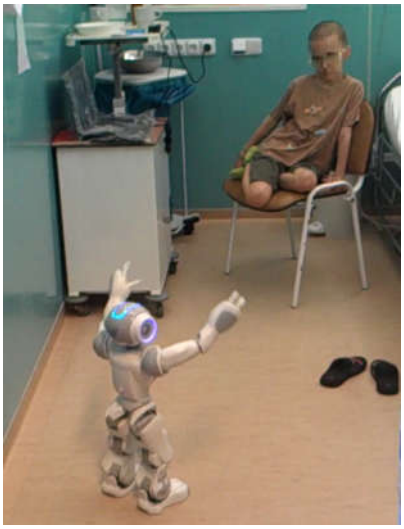
43. ábra. A NAO robot működési sémája [115]

A kutatás első fázisában négy 1-2 perces, egyszerű motivációs feladatot határoztunk meg a kórház pszichológusával és orvosaival együttműködve. Ezek egyrészt megtörhetik a monoton napi rutint, másrészt olyan feladatokat támogatnak, amelyeket a betegeknek naponta el kell(ene) végezni, de ezt gyakran nem (szívesen) teszik. A támogatott műveletek jellemzőit a 11. táblázat. tartalmazza (a Choregraphe NAO fejlesztői környezet dobozai és animációi).

11. táblázat. A kísérleti műveletek jellemző adatai [115]

Művelet	**Dobozok száma**	**Egyedi animációk száma**	**Időtartam (sec)**
Étkezés	53	10	102
Gyógyszer bevétel	25	5	72
Reggeli felkelés	20	6	75
Fürdés	16	2	60

A rendszeres étkezés alapvető fontosságú a gyógyuláshoz, de néhány gyógyszer csökkentheti az étvágya. Ebben NAO úgy tesz, mintha korogna a gyomra, elkezd enni, aztán még böfög is, ami nevetésre fakasztotta a gyermekeket. A gyógyszerek bevétele egy másik alapvető feltétele a gyógyulásnak, de sok gyerek lázad ellenük természetellenes megjelenésük és gyakran keserű ízük miatt. Ebben a jelenetben NAO úgy tesz, mintha elvesztek volna a gyógyszerek és igyekszik megtalálni őket. Kis hősökként írja le a szereket, amik a gonosz betegségek és vírusok ellen küzdenek. A keresés alatt NAO fel-le járkál, leguggol, kiemelve, hogy mennyire fontos minden egyes tabletta. A napi rutin része az időben történő felkelés és a reggeli torna. NAO azzal segít ebben, hogy köszönti a kis pácienseket és olyan egyszerű gyakorlatokat mutat be, amiket ők is könnyen utánozni tudnak. A fürdés is olyan feladat, amit a gyerekek gyakran utálnak. Különösen az, ha lavórban kell fürdeni. Ebben a jelenetben NAO fütyörészve sétál be, és azt imitálja, ahogy énekel zuhanyozás közben. A 44. ábra mutatja be a valós alkalmazási környezetet.



44. ábra. NAO akcióban
[115]

A legösszetettebb műveletek az étkezési jelenet, a legegyszerűbb pedig a fürdés. A megfelelő NAO mozdulatsorok meghatározása és programozása jelentős időt vett igénybe. A NAO robot meglehetősen korlátozott számítástechnikai erőforrásokkal rendelkezik, ezért az I. téziscsoport szerinti TTS rendszert implementáltuk rajta magyar nyelven. Így sikerült kellően érthető és valós idejű működést biztosítani. Az érthetőség különösen fontos, mert a környezet gyakran zajos és a NAO maximális hangereje az akkumulátor kímélése érdekében is mérsékelt.

A sterilitás biztosításához senki sem léphet be a betegszobába a megfelelő steril eljárások végrehajtása nélkül. Ezért cél az, hogy NAO vezérlése távolról is megoldható legyen. Ehhez járást és fordulást is végre kell hajtania bármilyen irányba. Ez erősen igénybe veszi a motorjait.

Kompenzálásként a leülés a megoldás. Ezzel együtt jár a felállás igénye is. Mivel a kórházban nem volt elérhető számítógépes környezet a NAO fejlesztését támogató Choregraphe szoftverrel, a jelenetek kiválasztását, elindítását és leállítását is infra (IR) távirányító segítségével oldottuk meg. Biztonsági funkciókra is szükség van. Ha NAO veszélyezteti saját magát, a beteget vagy bármilyen kórházi berendezést, le kell tudni állítani és visszaadni a vezérlést az infra távvezérlőnek. A NAO biztonsága érdekében egy esésdetektort is alkalmazunk. Ekkor minden motorja alapállapotba kerül és a vezérlést az IR eszköz veszi át.

A prototípust három, 4-14 év közötti leukémiás gyermekkel teszteltük. Koruk és hozzáállásuk is változó volt. A legelső probléma a teszt beillesztése volt a napi programjukba (reggeli vizit, tanulás vagy gyakorlás, stb.). A teszthez szülői hozzájárulást szereztünk be. Mielőtt NAO belépett a szobájukba, a robot felületét egy megfelelő folyadékkal kezelni kellett a fertőzés elkerülése érdekében. Az első tesztelő gyerek fiatal, gyors felfogású és nagyon aktív volt. Meg sem várta, hogy egy-egy jelenet befejeződjön, kérdésekkel közbevágott. Számára az interaktivitás hiánya akadályt jelentett, és noha tetszett neki a rendszer, túl sokáig nem tudta fenntartani a figyelmét. A második tesztelő a többieknél idősebb volt. Nem annyira a játék érdekelte a NAO kapcsán, hanem a műszaki megoldásai. A harmadik tesztelő a másik kettőnél jóval nagyobb szobában volt. Ezért a saját IR távvezérlőnk nem működött helyesen. De ez az akadály érdekes eredményt hozott. Rájött, hogy a TV-je távvezérlőjével tudja irányítani NAO-t. Megváltozott a szerep. Helyettünk a beteg vezérelte a robotot. A szoba mérete lehetővé tette, hogy hosszabban sétáljon és forgolódjon NAO. Egy másik esetben NAO elesett a csúszós padlón. Ez a helyzetet a beteg szerint is életszerűvé tette: ”Jó látni, hogy még NAO is hibázhat”. A szülőket is elvarázsolták NAO műszaki megoldásai és hálásak voltak, hogy szórakoztattuk gyermekeiket. Az orvosok is esélyesnek tartották bevonását a mindennapi életükbe, azonban további fejlesztéseket tartottak szükségesnek.

A legfontosabb kérdés a robot szerepének megtalálása a kórház életében. A gyerekek számára nem volt világos, hogy NAO-val játszhatnak, megérinthetik vagy csak egy költséges segédeszköz, amihez nem nyúlhatnak. Vajon orvosi kiegészítő, segédeszköz vagy egy segítő? Ezért fontos lenne a robotok helyét megtalálni a kórházi társadalomban. Ha NAO mindennapi társ lesz, akkor képesnek kell lennie a megújulásra, a változatosságra. A gyermekek különbözősége miatt a testre szabhatóság is elvárás. Összességében az alábbi következtetéseket vontuk le:

- a NAO robotot könnyen elfogadták a leukémiával küzdő gyerekek.
- a kétoldali interaktivitás (beszédfelismerés és megértés) rendszeres használathoz szükséges.

- A NAO abban az értelemben ideális, hogy viszonylag egyszerűen sterilizálható, kisméretű és a kísérletben vizsgált jelenetekhez megfelel az akkumulátor kapacitása.
- Az elesésre fel kell készíteni, fel kell tudnia állni és folytatni a jelenetet.
- A személyre szabhatóságot meg kell oldani.

Mindazonáltal a kísérleti projekt bizonyította, hogy egy humanoid robot sikeresen használható gyermekek csontvelő transzplantációs folyamata során, sajnos a prototípus elkészülése óta finanszírozás hiányában a továbblépés megakadt.

9.2.2. Gyógyszervonal

A Gyógyszervonal egy automatikus gyógyszer betegtájékoztató információs rendszer [119], [120], [121], [122]

Magyarországon a projekt megvalósításakor körülbelül ötezer törzskönyvezett gyógyszer volt (ami azóta csak növekedhetett), melyek engedélyezését az Országos Gyógyszerészeti Intézet (OGYI) végzi. Évente körülbelül 400 új gyógyszer jelenik meg és hozzávetőlegesen ugyanennyit vonnak ki a forgalomból. A projekt kitűzött célja, hogy elérhető legyen bárki számára hely- és időkorlát nélkül a gyógyszerekhez tartozó betegtájékoztató szövege. Az információs rendszer elsődleges célja, hogy telefonon keresztül elérhető legyen, és egy korszerű, automatikus beszédalapú dialógusrendszer segítségével a hívó fél számára felolvassa a kiválasztott gyógyszer betegtájékoztatóját. Ezenkívül WEB felületen is hozzáférhetők az adatok.

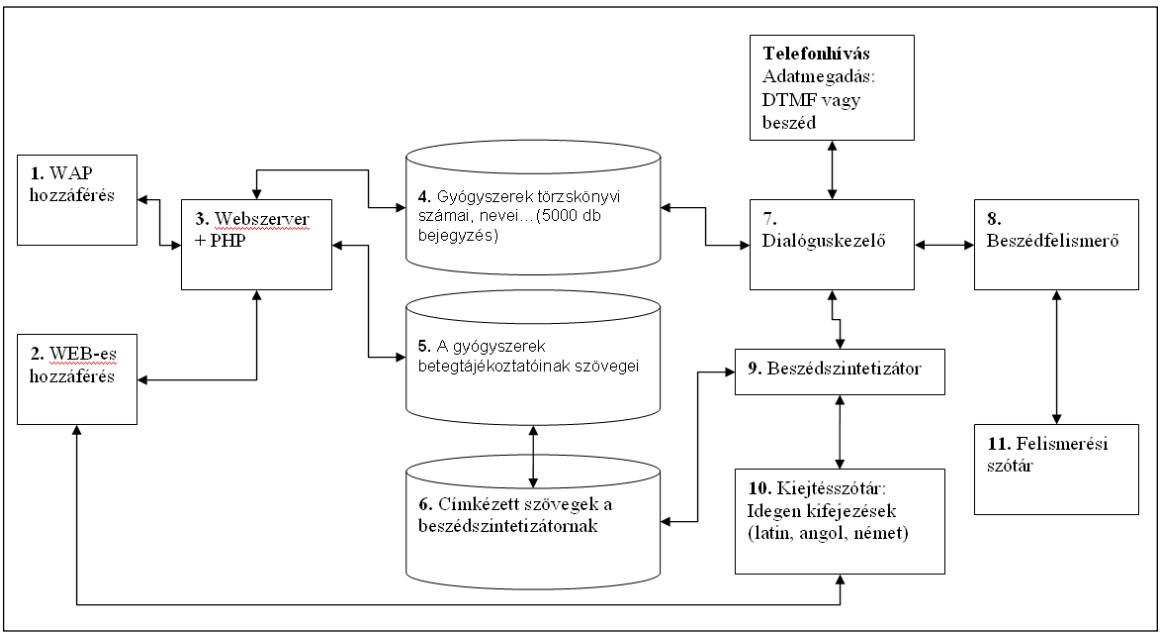
A legnagyobb célcsoport a legtöbb gyógyszert fogyasztók köre, vagyis az időskorúak (mintegy 3 millió nyugdíjas). A másik vélhető célcsoport a fiatalság, akik használják az Internetet. Ők segíthetnek az időseknek, ha megfelelően tájékoztatva vannak a szolgáltatás elérhetőségéről. Fontos célcsoport az orvosok köre is, akik az új gyógyszerekről ilyen módon is tájékozódhatnak. A szolgáltatásnak különös jelentősége van a vak és a látássérült emberek részére, mert ők a hagyományos, dobozba csomagolt tájékoztatót nem tudják elolvasni.

A rendszer főbb paraméterei a következők.

- 24 órás működés (bármikor hívható)
- többféle információs technológiával érhető el a gyógyszer-tájékoztató szövege (telefon, vezetékes és mobil Internet)
- a telefonvonal fogadó végén beszédfelismerő segíti az érdeklődőt, szóban kommunikálhat a géppel
- a gyógyszerismertetőt gép mondja el, így ezt akár többször is meg lehet hallgatni
- a gép precíz: megismételt hívás esetén ugyanazt az információt mondja el, ugyanabban a sorrendben, ugyanazon a hangon
- Internet használata esetén szövegben kapja meg az információt az ügyfél

- szakembereknek is tágabb teret ad a 24 órás szolgáltatási forma
- az információkérés nem hiúsul meg a vonal foglaltsága miatt (megfelelő számú csatorna üzemeltetése esetén)

A rendszer általános blokkvázlata a 45. ábra alapján tekinthető át.



45. ábra. A gyógyszerinformációs rendszer blokkvázlata [119]

A rendszerben minden adatot adatbázisban tárolunk, melyek konzisztenciájáért az Internetes modul szerkesztői része, illetve maga az üzembentartó a felelős. Az adatbázisban a gyógyszerek alapvető adatai (neve, törzskönyvi száma, hatáserőssége) mellett a hozzájuk tartozó betegtájékoztatók szövegeit, valamint a betegtájékoztatók szövegeinek szintetizált hullámformáit is tároljuk. Ezeket a szövegeket mondatonként tároljuk, minden mondatot csak egyszer.

Az előkészítő munka során meghatároztuk azon gyógyszerek listáját, amelyeket kezel a tájékoztató rendszer. Ezek a következők: vény nélküli gyógyszerek, vényre kapható nem kórházi felhasználású gyógyszerek. Minden gyógyszerhez tartozik egy törzskönyvi azonosító szám. Kialakítottuk a gyógyszernevek és a hozzájuk tartozó törzskönyvi-számok olyan adatbázisát, amely alapján a keresést el lehet végezni a rendszerben. A gyógyszerek forgalomba hozatalának engedélyezése hivatalos eljárás, az engedéllyel együtt kiadott alkalmazási előírás és betegtájékoztató hivatalos okiratnak számít. Biztosítani kellett azt, hogy a felolvasandó gyógyszertájékoztató szövege védve legyen az esetleges változtatásoktól, hiszen a rendszerben elektronikus formában, adatbázisokban tároljuk a szövegeket. A fejlesztés során szembekerültünk azzal a problémával is, hogy az eddigi betegtájékoztató jóváhagyási ügymenet során nem

figyeltek kellő mértékben a szöveg betű szintű helyességére. A gép a betűk szerint olvas, nem korrigál automatikusan, mint a szem, amikor emberek olvasnak. Az elütések, a helyesírási hibák rontják a beszédszintetizátor érthetőségét. A következő főbb hibacsoportokat állapítottuk meg.

- Betűkimaradás: mértétől /mértékétől/; időponját /időpontját/; h a gyógyszer/ha a gyógyszer/; zolgál /szolgál/
- Betűbetoldás: magzatatot /magzatot/; Aeurius /Aerius/;
- Betűcsere : Bleocin inejkció/Belocin injekció/
- Helytelen karakter a szövegben (elütés): 1 x ″ 4 mg-os tabletta
- Helytelen karakter konverzió: legfeljebb 25^0C-on/25°C-on/; 25şC-on/25°C-on/
- Rövidítés helytelen írásmódja, nincs pont utána: ill ;
- Mondat a mondatban - ...kevés folyadékkal (nem grapefruit lével!) étkezés után...
- Idegen szavak többféle írásmódja – migraine és migrain
- Nem egységes szövegszerkezet: más a logikai sorrend, mivel minden gyár másfajta fogalmazást valósít meg

Ezeket a hibákat javítani kell. A korrigálásra olyan korrektúrázó eljárást fejlesztettünk ki, amelyik nem sérti a hivatalos okirattal szemben támasztott követelményeket.

A felhasználói felületek közül a legbonyolultabb a telefonos rendszer működtetését biztosító dialógus („párbeszéd” az ügyfél és a gép között) megtervezése és kialakítása. Ebben biztosítani kell az ember – gép közötti élő párbeszéd optimalizált, mégis kötött formáját. A tervezéskor a legnagyobb problémát a gyógyszerek keresésének, azonosításának egyszerű megvalósítása jelenti (a hívó fél szeretné egy gyógyszerismertető felolvasását kérni, ehhez a gépnek meg kell azt találni a belső adatbázisokban). A gyógyszereket a dialógusban alapvetően két különböző módszerrel azonosíthatjuk, vagy a telefon billentyűzetével bevisszük az gyógyszer valamelyik egyedi adatát, vagy bemondjuk a gyógyszer nevét, amelyből a beszédfelismerő megpróbálja azonosítani a gyógyszert az adatbázisban.

A nyomógombos bevitelnél több lehetőség közül választhat a tervező. Az egyik kézenfekvő megoldás, amikor a rendszer a gyógyszer ún. törzsszámának bebillentyűzését kéri a hívó féltől. A gyógyszer törzsszáma egy rövid azonosító, amely egy szöveges résszel kezdődik, majd egy általában 4-5 karakteres számmal fejeződik be. A szöveges rész nem lényeges, a 4-5 karakteres számot könnyű bebillentyűzni a nyomógombokkal. Ez biztos eredményt ad, de az emberek többsége nem ismeri ezeket a számokat, a gyógyszer dobozán sem találhatók meg egyértelműen, valamint a vakok és gyengén látók ezt el sem tudják olvasni, amíg Braille-írással fel nem tüntetik. Egy alternatív megoldás lehet, hogy a felhasználó bebillentyűzheti a gyógyszer nevét is, hasonlóan az SMS íráshoz. Ez főleg idősebb felhasználóknál nem lehet népszerű. A harmadik eset, hogy az ABC betűcsoportjaihoz gombokat rendelünk, például: ABCD=1-es gomb, EFGH=2-es gomb, hasonlóan a telefonos prediktív (T9) bevitelhez. A gyógyszer nevének betűit

szóban kéri be a rendszer (például: adja meg a gyógyszer első betűjét a megfelelő gomb megnyomásával). Átlagosan 10-12 billentyűnyomással meghatározható a keresett készítmény.

A bemondáson alapuló megoldáshoz **beszédfelismerő**t kell alkalmazni. Ez természetesebb a felhasználó számára, azonban rendszertechnikailag sok új problémát vet fel. A legnagyobb probléma a gyógyszernevek természetéből adódik, mivel ezek általában latin alapú elnevezések, amelyeknek nincs széles körben elfogadott, egységes kiejtése, emellett esetleg több szóból is állhatnak. Az ügyfélnek a gyógyszer nevét kell bemondania a telefonba és a beszédfelismerő azonosítja azt a belső felismerési szótára segítségével. Ez sem egyértelmű, hiszen a gyógyszer neve mellett gyakran szerepel a gyártó neve is (például: Bayer Aspirin), vagy valamilyen hatáserősségre utaló szám (Vitamin C 100 mg filmtabletta). Előre nem lehet tudni, hogy a hívó fél hogyan fogja mondani. A gyógyszer nevének kiejtési variáltsága is többféle lehet. Fel kell mérni azt, hogy mi lehet az emberektől elvárható kiejtés és több variációra is fel kell készülni.

A fent leírt módszerekkel sok esetben a gép nem tudja egyértelműen azonosítani a gyógyszert, több jelöltet is talál az adatbázisban. A tervezési célunk az, hogy 3-5 lehetséges készítményre lehessen leszűkíteni a keresés eredményét. Ekkor már lehetőség van a készítmények egyenkénti felsorolására, amelyből a felhasználó már kiválasztja azt, amelyikre gondolt. A gyógyszer kiválasztása után a rendszer felajánlja az ügyfélnek, hogy az adott betegtájékoztató melyik fejezetét (ld. 12. táblázat) akarja hallani.

Az adott fejezeten belül, – miután a gép elkezdte a felolvasást – lehetőség van a mondatok között előre, hátra ugrani, illetve az aktuális mondatot megismételtetni. Az internetes felületeknél a kiválasztás és a megjelenítés megvalósítása egyszerűbb, mivel itt billentyűzeten és kijelzőn keresztül történik a kommunikáció az ügyfél és a gépi rendszer között. Akár egy keresőszóra a rendszer által talált 3-5 jelölt közül a felhasználó ki tudja választani a képernyőn, hogy melyik gyógyszerről kéri a tájékoztatót.

A gyógyszerinformációs rendszer beszédszintetizátora az I. téziscsoport eredményei szerinti Profivox szövegfelolvasóra alapozott speciális szoftver, amelyik kifejezetten erre a célfeladatra készült (Profivox-Med). A szoftver specialitását két pontban lehet összegezni. Az egyik, hogy érzékeli a latin és egyéb idegen nyelvű szakszavak jelenlétét a szövegben és azokat a magyar kiejtésnek megfelelően olvassa fel. A másik, hogy fel van készítve a gyógyszerészek által használt speciális nyelvezet (mondatszerkesztés, szóhasználat) mondatprozódiai értelmezésére, feldolgozására és megvalósítására.

12. táblázat. A betegtájékoztató hat fejezete [119]

Az eredeti sablon szerint a fejezet címe a doc fájlban	Jelző karaktersorozattal ellátva, gépi szortírozáshoz, kereséshez
1. Milyen típusú gyógyszer X és milyen betegségek esetén alkalmazható?	<<<1>>> 1. Milyen típusú gyógyszer a/az X és milyen betegségek esetén alkalmazható?
2. Tudnivalók az X <szedése> <alkalmazása> előtt	<<<2>>> 2. Tudnivalók az X <szedése> <alkalmazása> előtt
3. Hogyan kell <szedni> <alkalmazni> X-t	<<<3>>> 3. Hogyan kell <szedni> <alkalmazni> X-t
4. Lehetséges mellékhatások	<<<4>>> 4. Lehetséges mellékhatások
5. A készítmény tárolása	<<<5>>> 5. A készítmény tárolása
6. További információk	<<<6>>> 6. További információk

A Gyógyszervonal beszédalapú telefonos felhasználói felületéhez egy gyógyszernévre optimalizált, nagyméretű kötött szótárból dolgozó beszédfelismerő is tartozik. A felismerő szoftver a felhasználó által a telefonba bemondott gyógyszer nevét ismeri fel, és így azonosítja azt a belső adatbázisban. Az eredményt közli a felhasználóval. A felismerő az alábbi tulajdonságokkal rendelkezik:

- telefonon bemondott gyógyszernevek felismerése elfogadható (min. 90%) pontossággal,
- új gyógyszerek megjelenése esetén az egyszerű bővíthetőség biztosítása,
- a telefonos felhasználói felület menürendszerében való navigáláshoz szükséges parancsszavak és opciók felismerése nagy pontossággal.

A beszédfelismerő motor beszélőfüggetlen, nyílt szótárral rendelkezik, azaz elvileg tetszőleges szó felismerésére képes (a szó meghatározása után)]. Elvileg, mert:

- a szavakat helyesen (a kiejtésnek megfelelően) kell megadni a rendszernek,
- ügyelni kell, hogy nagyon hasonló szavak ne kerüljenek a rendszerbe,
- ha mégis vannak hasonló szavak, dialógus szinten fel kell tudni készülni a tévesztési lehetőségekre,
- a sok gyógyszernév miatt a valós idejű feldolgozás speciális megfontolásokat igényel.

A „gyógyszervonal" tájékoztató rendszer az első automatikus, nyilvános informatikai tudakozó volt a gyógyszerek betegtájékoztatójának szövegével kapcsolatosan Magyarországon. A rendszert a BME TMIT és az OGYI közösen fejlesztette a GVOP pályázati támogatási rendszer keretében. A „Gyógyszervonal" szolgáltatást az OGYI vezette be és üzemeltette 2007. januárjától addig, amíg az átszervezések folyamán a támogatása abbamaradt. A rendszert az Európai Bizottság

szakértője a regionális támogatásokból megvalósult innovatív szolgáltatások közé sorolta [122], [123].

9.3. Fogyatékos és idős embereket támogató szolgáltatások

Terjedelmi korlátok miatt az ebben és következő témakörben született alkalmazásainkat csak felsorolom. A rendszerek különböző változatai mind a négy téziscsoport eredményeire építenek. Kiemelem, hogy a VoxAid alkalmazás sztrókos betegek rehabilitációjára optimalizált változata az EIT Digital európai startup ötlet versenyén III. helyezést ért el [123].

- a PAELIFE EU AAL projekt keretében idős emberek infokommunikációs szolgáltatásainak támogatására [124][4],
- VoxAid2006 prototípus siketnéma emberek telefonálásának támogatására [99]
- VoxAid2012 prototípus beszédsérült emberek mindennapi kommunikációjának támogatására, logopédiai és afáziás betegek rehabilitációjának támogatására [100], [123]
- Jaws for Windows képernyő olvasó program, ami a legelterjedtebb PC-s hasonló eszköz Magyarországon (2000-től több változat folyamatos fejlesztés alatt),
- RoboBraille (www.robobraille.org), [125] többnyelvű gépi szöveg fájl - beszéd fájl átalakító ingyenes internetes szolgáltatás kiegészítése magyar nyelvre (2012-),
- beszélő mobil alkalmazások vak emberek számára Symbian, Windows Phone és Android platformon [126] [127],
- a VUK EU AAL projekt keretében látássérült emberek beltéri navigációjának támogatására (2019),

9.4. Általános információs rendszerek

- a www.metnet.hu időjárás portál, illetve a Microsoft 2013-as fejlesztői versenyén nyertes *Időjárás Mindenkinek* Windows8 alkalmazás,
- egy távközlési szolgáltató automatizáltan kialakított interaktív hangválasz (IVR) rendszere (2009-től),
- beszéd-dialógus mintarendszer intelligens lakás prototípusban a BelAmi projekt keretében (2007),
- beszédvezérelt okosTV készülék prototípus (2014),
- Szlovén-magyar hangos szótár (2018),
- www.webforditas.hu többnyelvű internetes fordító szolgáltatás (2006-, a Google Translate-et 2 évvel megelőzve).

[4] https://www.youtube.com/watch?time_continue=9&v=ads85G3ArZI

10. A tézisek összefoglalása egységes szerkezetben

I. **téziscsoport: A d**iád és triád elemek összefűzésén alapuló gépi szövegfelolvasás

I.1 tézis: A diád és triád elemösszefűzéses gépi szövegfelolvasó eljárás
Kidolgoztam a magyar nyelv sajátosságainak megfelelő első diád és triád hullámforma elemösszefűzéses gépi szövegfelolvasó eljárás rendszertervét (ld. 6. ábra), amely diád és triád méretű magyar hangkapcsolódások felhasználásával készít gépi beszédet, és igazoltam, hogy az ezek felhasználásával létrehozott rendszer MOS (Mean Opinion Score) szubjektív értékelés szerint jobb hangminőséget ad, mint a korábbi, más elven működő megoldások (például Hungarovox [28], Brailab [44], PC talker [45]) Az eljárást kiterjesztettem német nyelvre is.

I.2. tézis: Diád és triád alapú rendszerek beszédadatbázisa
Megterveztem az első magyar diád és triád hullámforma elemek megvalósításához felhasználható magyar nyelvű felolvasásos beszédadatbázis szerkezetét és az annak elkészítéséhez szükséges, az átlagos prozódiai jellemzőket biztosító szövegkorpuszt.

II. téziscsoport: Célorientált, korpusz-alapú gépi felolvasó rendszerek

II.1. tézis: Magyar nyelvű korpusz-alapú gépi szövegfelolvasás modellje
Kidolgoztam magyar nyelvre az első korpusz-alapú hangnyomás-idő függvények automatikus válogatásán alapuló gépi szövegfelolvasó eljárás modelljét, amely szavak, szókapcsolatok, mondatrészek hangnyomás-idő függvényeinek célorientált összefűzésével készít gépi beszédet, valamint az ehhez kapcsolódó, fonetikai szempontok szerint kialakított költségfüggyényeket és indirekt prozódiai modellt. MOS vizsgálatokkal igazoltam, hogy jobb hangminőséget eredményez, mint az I. téziscsoport szerinti megoldások.

II.2. tézis: A korpusz-alapú szövegfelolvasó tématerületekhez történő adaptálása
Egységes eljárást és többszintű modellt dolgoztam ki elsőként a korpusz-alapú hullámforma elemválogatáson alapuló magyar nyelvű szövegfelolvasó technológia különböző tématerületekhez illetve több- vagy kevert nyelvű alkalmazáshoz történő adaptálására. A megoldás működőképességét, valamint az emberi felolvasással való összetéveszthetőségét három (időjárás-jelentés, pályaudvari hangos információ szolgáltatás és árlista-felolvasás) különböző tématerületen igazoltam.

II.3. tézis: A gépi szövegfelolvasás prozódiai változatosságának megvalósítása
Új módszert dolgoztam ki prozódiai frázisok hasonlósága alapján képzett prozódiai csoportok létrehozásához és ezekből nem determinisztikus válogatással gépi szövegfelolvasó rendszerek prozódiai változatosságát tettem lehetővé. Megmutattam, hogy egy magyar nyelvű megvalósítás során a felhasználók ezt a módszert a hagyományos szabály-alapú és a II.1-es tézis szerinti indirekt megoldásnál is jobbnak értékelték. Ez a prozódiai modell alkalmazható a hagyományos elemösszefűzéses, a korpusz-alapú és a HMM rendszerekben egyaránt.

III. téziscsoport: Statisztikus parametrikus gépi szövegfelolvasó rendszerek

III.1 Tézis: A rejtett Markov modell alapú magyar nyelvű gépi felolvasó rendszer
Azonosítottam az újonnan megalkotandó vagy adaptálandó rendszermodulokat az első gépi tanuláson alapuló magyar nyelvű gépi szövegfelolvasó rendszer kialakításához. Létrehoztam egy olyan adatstruktúra modellt, ami alapján az ezen az elven alapuló gépi szövegfelolvasó rendszer hatékonyan megvalósítható.

III.2 Tézis: A HMM TTS rendszer minőségének javítása
Új elven, a maradékjelre alkalmazott elemkiválasztásos eljáráson alapuló, megvalósítást elősegítő koncepciót és modellt alkottam a HMM TTS rendszerben alkalmazandó jobb minőségű beszédkódolók létrehozásához.

III.3. Tézis: Rövid és kérdő mondatok jobb minőségű megvalósítása
Kidolgoztam a magyar kérdő mondatok alapfrekvencia-idő függvényeinek statisztikai modellezését gépi beszédelőállításhoz.

IV. téziscsoport: Multimodális beszédinformációs rendszerek

IV.1. tézis: Mobil felhasználói felületek modalitásainak szinkronizálása
Új, skálázható, multimodális leíró nyelvet alkalmazó eljárást dolgoztam ki mobil multimodális felhasználói felületek modalitásainak szinkronizálására. A módszer működőképességét a grafikus és a beszéd modalitás szinkronizálását megvalósító mintaalkalmazásokkal igazoltam.

IV.2. tézis: Kommunikációs kontextust jelző akusztikus jelkészlet előállítása
Kidolgoztam kommunikációs kontextust jelző új akusztikus jelkészlet (spemoticon-ok) elméletét és modelljét, valamint annak megvalósítási módszerét gépi szövegfelolvasó eszközrendszerére alapozva. Megalkottam egyfajta jelkészlet csoportot. Objektív paraméterbeállítások módszerével és szubjektív tesztekkel igazoltam a módszer eredményességét.

IV.3. tézis: Multimodális felhasználói felületek beszédsérült emberek támogatására
Új módszert dolgoztam ki multimodális felhasználói felületek hatékony felhasználására beszédsérült emberek kommunikációjának támogatására. A módszert a gépi szövegfelolvasó rendszerekben többféle szövegbeviteli formára és eszközplatformra (asztali számítógép, notebook, okostelefon, tablet) alkalmaztam.

Köszönetnyilvánítás

Köszönöm elsősorban a BME TMIT Beszédkommunikáció és Intelligens Interakciók Laborcsoport korábbi és mai tagjainak (Gordos Géza, Olaszy Gábor, Olaszi Péter, Kiss Géza, Zainkó Csaba, Bőhm Tamás, Gyires-Tóth Bálint, Csapó Tamás, Bartalis Mátyás, Laczkó Klára, Nagy Péter, Mohammed Al-Radhi, Sevinj Yolchuyeva, Hajgató Gergely, Moni Róbert, Hamdi Abed, Mihajlik Péter, Fegyó Tibor, Tarján Balázs, Vicsi Klára, Szaszák György, Sztahó Dávid, Kiss Gábor, Tulics Miklós) csapatmunkáját, másrészt a BME TMIT vezetőinek, munkatársainak, hallgatóimnak és kutatási partnereinknek az együttműködését, ami a jelen dolgozatban bemutatott eredményeimet is lehetővé tette. Sokat javítottak az értekezés színvonalán Imre Sándor, Olaszy Gábor és Sallai Gyula értékes megjegyzései, ezt külön köszönöm nekik.

Terjedelmi korlátok miatt csak néhány, több évtizedes intézményi együttműködést tudok felsorolni: MTA Nyelvtudományi Intézet, ELTE Fonetika Tanszék, Szegedi Tudományegyetem Mesterséges Intelligencia Kutatócsoport, MTA SZTAKI, MTA Természettudományi Kutatóközpont, Magyar Telekom, IT.DOT Kft, Morphologic Kft, Informatika a Látássérültekért Alapítvány, Bay Zoltán Alkalmazott Kutatási Közhasznú Nonprofit Kft.

A téziseimben áttekintett kutatások eredményei többek között a BelAmi, GVOP 3.1.1-2004-05-0426, TÁMOP-4.2.1/B-09/1/KMR-2010-0002, CESAR (ICT PSP No 271022, EU_BONUS_12-1-2012-0005,), PAELIFE (AAL_08-1-2011-0001), VUK (AAL-2014-1-183), DANSPLAT (Eureka 9944) valamint az EITKIC_12-1-2012-0001 projekt keretében jöttek létre (a projektek a Kutatási és Technológiai Innovációs Alap valamint az Európai Bizottság támogatásával valósultak meg).

Irodalomjegyzék

[1] G. Németh és G. Olaszy, szerk., A magyar beszéd, Budapest: Akadémiai Kiadó, 2010, p. 749.

[2] K. N. Stevens, S. Kasowski és C. G. M. Fant, „An electrical analog of the vocal tract," *Journal of the Acoustical Society of America vol. 24. issue 2,* p. 734–742, 1953.

[3] G. Olaszy, Elektronikus beszédelőállítás. A magyar beszéd akusztikája és formánsszintézise., Budapest: Műszaki Könyvkiadó, 1989.

[4] D. H. Klatt és L. C. Klatt, „Analysis, synthesis, and perception of voice quality variations among female and male talkers," *The Journal of the Acoustical Society of America vol. 87., issue 2,* pp. 820-857, 1990.

[5] A. E. Rosenberg, R. W. Schafer és L. R. Rabiner, „Effects of Smoothing and Quantizing the Parameters of Formant-Coded Voiced Speech," *J. Acoust. Soc. Am.,* pp. Volume 50, Issue 6B, pp. 1532-1538, 1971.

[6] E. Moulines és F. Charpentier, „Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communications 9.,* p. 453–467, 1990.

[7] M. Beutnagel, A. Conkie, S. J. Y. Stylianou és A. Syrdal, „The AT&T next-gen TTS system," *Journal of the Acoustical Society of America, Vol. 105, Issue 2,* 1999.

[8] G. Olaszy, G. Németh, P. Olaszi, G. Kiss, C. Zainkó és G. Gordos, „Profivox – a Hungarian TTS System for Telecommunications Applications," *International Journal of Speech Technology. Vol 3-4.,* pp. 201-215, 2000.

[9] G. Németh, G. Olaszy és M. Fék, „Új rendszerű, korpusz alapú gépi szövegfelolvasó fejlesztése és kísérleti eredményei," in *Beszédkutatás 2006*, Budapest, 2006, pp. 183-196.

[10] H. Zen, K. Tokuda és A. W. Black, „Statistical parametric speech synthesis," *Speech Communication, vol. 51,* pp. 1039-1064, 2009.

[11] H. Zen, A. Senior és M. Schuster, „Statistical Parametric Speech Synthesis Using Deep Neural Networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New York, 2013.

[12] P. Nagy és G. Németh, „DNN-Based Duration Modeling for Synthesizing Short Sentences," in *Speech and Computer : 18th International Conference*, Budapest, 2016.

[13] S. Yolchuyeva, G. Németh és B. Gyires-Tóth, „Text normalization with convolutional neural networks," *International Journal of Speech Technology, Vol. 21, Issue 3,* p. 589–600, 2018.

[14] M. Mori, „The uncanny valley," *(K. F. MacDorman & N. Kageki, Trans.). IEEE Robotics & Automation Magazine, Vol. 19 Issue 2,* p. 98–100, 1970/2012.

[15] P. Olaszi, Magyar nyelvű szöveg-beszéd átalakítás: nyelvi modellek, algoritmusok és megvalósításuk, PhD disszertáció: BME, 2002.

[16] T. M. Bőhm, Analysis and modeling of speech produced with irregular phonation, PhD Dissertation: BME, 2009.

[17] C. Zainkó, Gépi beszédkeltés infokommunikációs rendszerekben, PhD disszertáció: BME, 2010.

[18] T. G. Csapó, A gépi beszéd-előállítás természetességének növelése, PhD disszertáció: BME TMIT, 2013.

[19] B. Tóth, Rejtett Markov-modell alapú gépi beszédkeltés, PhD disszertáció: BME TMIT, 2013.

[20] G. Németh, „Kempelentől a WaveNet-ig: a gépi beszédkeltés tudományának fejlődése," in *A humán tudományok és a gépi intelligencia*, G. Tocsvai Nagy, Szerk., Budapest, Gondolat Kiadó, 2018, pp. 127-155.

[21] F. Kempelen, Az emberi beszéd mechanizmusa, valamint a szerző beszélőgépének leírása, Budapest: Szépirodalmi Könyvkiadó, 1989.

[22] M. Bánó, „Tetszőleges szöveg reprodukálására alkalmas beszélőgép". Magyarország Szabadalom száma: 74361 , 30 11 1916.

[23] H. Dudley, R. R. Riesz és S. A. Watkins, „A Synthetic Speaker," *J. Franklin Inst. 227,* pp. 739-764. (Reprinted in Flanagan and Rabiner, 1973), 1939.

[24] F. Cooper, „Speech synthesizers," in *The Hague: Mouton & Co*, Helsinki, 1961.

[25] G. Olaszy, „Szintetizált magyar magánhangzók formáns-intenzitás és formáns-sávszélesség értékei.," *Magyar fonetikai füzetek,* pp. 68-77, 1978.

[26] G. Gordos és G. Takács, Digitális beszédfeldolgozás, Budapest: Műszaki Könyvkiadó, 1983.

[27] P. Mermelstein, „Articulatory model for the study of speech production," *Journal of the Acoustical Society of America 53 (4),* pp. 1070-1082, 1973.

[28] G. Kiss és G. Olaszy, „A Hungarovox magyar nyelvű, szótár nélküli, valós idejű párbeszédes beszédszintetizáló rendszer," *INFORMÁCIÓ ELEKTRONIKA, Vol. 19/2,* pp. 98-111, 1984.

[29] D. Klatt, „How Klattalk became DECtalk: An Academic's Experiences in the Business World," in *The Official Proceedings of Speech Tech '87*, New York, 1987.

[30] B. Möbius, „Corpus-based speech synthesis: methods and challenges," in *Speech and Signals - Aspects of Speech Synthesis and Automatic Speech Recognition*, W. F. Sendlmeier és W. Hess, szerk., Frankfurt am Main, Hector, 2000, p. 79–96.

[31] C. J. Plomp és O. Mayora-Ibarra, „A generic widget vocabulary for the generation of graphical and speech-driven user interfaces," *International Journal of Speech Technology,* pp. 39-47., 2002.

[32] J. L. Dvorak, „Method and system for unified speech and graphic user interfaces". Washington, DC: U.S. Patent and Trademark Office. Szabadalom száma: 7,389,235., 2008.

[33] C. Zainkó, M. Bartalis, G. Németh és G. Olaszy, „A Polyglot Domain Optimised Text-To-Speech System for Railway Station Announcements," in *INTERSPEECH 2015*, Dresden, 2015.

[34] G. Kiss, G. Németh, G. Olaszy és G. Gordos, „A Flexible Multilingual TTS Development and Speech Research Tool," in *International Conference on Speech Communication and Technology (Interspeech 2001)*, Aalborg, Denmark, 2001.

[35] G. Olaszy és G. Németh, „IVR for Banking and Residential Telephone Subscribers Using Stored Messages Combined with a New Number-to-Speech Synthesis Method.," in *Human Factors and Voice Interactive Systems.*, New York, Kluwer Academic Publishers, 1999, pp. 237-256.

[36] M. Gósy, „BEA – A multifunctional Hungarian spoken language database," *PHONETICIAN Vol. 105/10,* pp. 50-61, 2013.

[37] P. Mihajlik, T. Fegyó, Z. Tüske és P. Ircing, „A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages - like Hungarian," *Proc. of Interspeech,* pp. 1497-1500, 2007.

[38] P. Boersma és D. Weenink, „Praat: doing phonetics by computer [Computer," 2012. [Online]. Available: http://www.praat.org/. [Hozzáférés dátuma: 09 03 2012].

[39] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi és T. Kitamura, „Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. of ICASSP*, Istanbul, Turkey, 2000.

[40] Z. Wu és O. W. a. S. King, „Merlin: An Open Source Neural Network Speech Synthesis System," in *Proceedings of the 9th ISCA Speech Synthesis Workshop*, Sunnyvale, USA, 2016.

[41] F. Chollet, *Keras: Theano-based deep learning library,* Code: https://github.com/fchollet. Documentation: http://keras.io., 2016.

[42] M. Abadi és é. tsai, *TensorFlow: Large-scale machine learning on heterogeneous systems,* Software available from tensorflow.org.: tensorflow.org., 2015.

[43] ITU-R Recommendation BS.1534, *Method for the subjective assessment of intermediate audio quality,* 2001..

[44] A. Arató, *A BraiLab beszélő számítógépcsalád,* Budapest: Kandidátusi értekezés, 1984.

[45] J. Király, „A PC-TALKER beszédszintetizátor és digitális hangrögzítő-visszajátszó rendszer," *Magyar Elektronika,* %1. kötet6. évf. , %1. szám12. szám, 1989.

[46] G. Bailly és C. S. T. Benoit, szerk., Talking Machines: Theories, Models, and Designs, Amsterdam: Elsevier Science & Technology Books, 1992.

[47] R. W. Sproat és J. P. Olive, „Text-to-speech synthesis," *AT&T technical journal, Vol. 74, issue2,* pp. 35 - 44, 1995.

[48] G. Németh, C. Zainkó, L. Fekete, G. Olaszy, G. Endrédi., P. Olaszi, G. Kiss and P. Kis, "The design, implementation and operation of a Hungarian e-mail reader," *INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY,* Vols. 3:(3-4), pp. 217-236, 2000.

[49] E. Straub, „MATÁV 1999-es éves jelentés," MATÁV, Budapest, 2000.

[50] J. Koch, H. Jung, J. Wettach, G. Nemeth, K. Berns, S. Lee és S. Mun, „Dynamic speech interaction for robotic agents," in *Recent Progress in Robotics: Viable Robotic Service to Human*, S. Lee, S. I. és K. M. , szerk., Berlin, Heidelberg, Springer, 2008, pp. 303-315.

[51] G. Németh, G. Olaszy, M. Bartalis, C. Zainkó, M. Fék és P. Mihajlik, „Beszédadatbázisok előkészítése kutatási és fejlesztési célok hatékonyabb támogatására," *HIRADÁSTECHNIKA,* pp. LXIII:(5) pp. 18-24, 2008.

[52] G. Olaszy, G. Németh és G. Gordos, „The MULTIVOX multilingual text-to-speech converter," in *Talking machines: Theories, Models and Applications*, Amsterdam, North-Holland Publishing Company, 1992, pp. 385-411..

[53] G. Olaszy, G. Kiss és G. Németh, „Hungarian audiovisual prosody composer and TTS development environment," in *Prosody 2000 (szerk. Puppel S, Demenko G)*, Poznan, Adam Mickiewicz University, 2001b, pp. 167-177.

[54] E. Straub, *MATÁV 2003-as éves jelentés,* Budapest: MATÁV, 2004.

[55] G. Németh, G. Kiss, C. Zainkó, G. Olaszy és B. Tóth, „Speech Generation in Mobile Phones," in *Human Factors and Interactive Voice Response Systems*, New York, Springer, 2008, pp. 63-191.

[56] Z. NISZ, „Akadálymentes Magyarország," 2018. [Online]. Available: http://akadalymentes.magyarorszag.hu/. [Hozzáférés dátuma: 06 08 2018].

[57] A. Nagy, P. Pesti, G. Németh és T. Bőhm, „Design issues of a corpus-based speech synthesizer," . *HÍRADÁSTECHNIKA, LX:(6),* pp. 6-12., 2005.

[58] G. Németh, C. Zainkó, M. Bartalis, G. Olaszy és G. Kiss, „Human Voice or Prompt Generation? Can They Co-Exist in an Application?," in *Interspeech 2009*, 2009.

[59] M. Fék, P. Pesti, G. Németh és C. Zainkó, „Generációváltás a beszédszintézisben," *HÍRADÁSTECHNIKA LXI:(3),* pp. 21-30, 2006.

[60] H. Kawai, T. Toda, J. Ni, T. M és K. Tokuda, „Ximera: a new TTS from ATR based on corpus-based technologies," in *Proc. of the 5th ISCA Speech Synthesis Workshop*, Pittsburgh, 2004.

[61] G. Olaszy, *A beszéd akusztikai-fonetikai elemzése és modellezése különös tekintettel a korszerű beszédépítés követelményeire,* MTA Doktora disszertáció, 2001.

[62] G. Olaszy, „Az artikuláció akusztikai vetülete - a hangsebészet elmélete és gyakorlata.," in *Kísérleti Fonetika Laboratóriumi Fonológia a Gyakorlatban (KIFLAF).*, Debrecen, Debreceni Egyetem Kossuth Kiadója, 2003, pp. 241-254.

[63] A. Viterbi, „Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *EEE Transactions on Information Theory, Vol. 13, no. 2,,* pp. 260-269, April 1967.

[64] L. R. Cormen, „Chapter 17 "Greedy Algorithms"," in *Introduction to Algorithms*, Mcgraw-Hill, 1990, p. 768.

[65] Hungarobyte, Kft, „Digiton rendszerek," 1989-2019. [Online]. Available: http://www.hungarobyte.hu/hungarobyte.php. [Hozzáférés dátuma: 20 06 2019.].

[66] G. Németh, M. Fék és T. Csapó, „Increasing Prosodic Variability of Text-To-Speech Synthesizers," in *Interspeech 2007*, 2007.

[67] T. G. Csapó, C. Zainkó és G. Németh, „A Study of Prosodic Variability Methods in a Corpus-Based Unit Selection Text-To-Speech System," *INFOCOMMUNICATIONS JOURNAL, LXV:(1),* pp. 32-37, 2010.

[68] T. G. Csapó és G. Németh, „Prozódiai változatosság rejtett Markov-modell alapú szövegfelolvasóval," in *VIII. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, 2012.

[69] M. Chu, Y. Zhao és E. Chang, „Modeling stylized invariance and local variability of prosody in text-to-speech synthesis," *Speech Communication, Vol. 48.,* p. 716–726, 2006.

[70] B. P. Tóth, B. Szórádi és G. Németh, „Improvements to Prosodic Variation in Long Short-Term Memory Based Intonation Models Using Random Forest," in *SPECOM 2016*, Budapest, 2016.

[71] A. A. Markov, „An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains.," *Bulletin of the Imperial Academy of Sciences of St. Petersburg,* pp. 153-162, 1913.

[72] F. Jelinek, „Continuous speech recognition by statistical methods," *Proc. IEEE, vol. 64,* pp. 532-536, 1976.

[73] B. P. Tóth és G. Németh, „Hidden Markov Model Based Speech Synthesis System in Hungarian," *INFOCOMMUNICATIONS JOURNAL, LXIII:(7),* pp. 30-34, 2008.

[74] B. P. Tóth és G. Németh, „Rejtett Markov-modell alkalmazása magyar nyelvű gépi szövegfelolvasóhoz," *BESZÉDKUTATÁS, 16,* pp. 182-193., 2008.

[75] K. Vicsi, A. Kocsor, C. Teleki és L. Tóth, „Beszédadatbázis irodai számítógép-felhasználói környezetben," in *II. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, 2004.

[76] T. G. Csapó és G. Németh, „A novel codebook-based excitation model for use in speech synthesis," in *IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, 2012.

[77] H. Kawahara, I. Masuda-Katsuse és A. de Cheveigné, „Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneousfrequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication, vol. 27, no. 3,* p. 187–207, 1999.

[78] K. Koishida, K. Tokuda, T. Kobayashi és S. Imai, „Spectral representation of speech using mel-generalized cepstral coefficients," *The Journal of the Acoustical Society of America, 100 (4),* 1996.

[79] T. G. Csapó és G. Németh, „Modeling irregular voice in statistical parametric speech synthesis with residual codebook based excitation," *IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING, 8:(2),* pp. 209-220, 2014a.

[80] T. G. Csapó és G. Németh, „Statistical parametric speech synthesis with a novel codebook-based excitation model," *INTELLIGENT DECISION TECHNOLOGIES, 8:(4),* pp. 289-299, 2014b.

[81] T. G. Csapó és G. Németh, „Automatic transformation of irregular to regular voice by residual analysis and synthesis," in *Interspeech 2015*, Dresden, 2015.

[82] T. G. Csapó, G. Németh, M. Cernak és P. N. Garner, „Modeling Unvoiced Sounds In Statistical Parametric Speech Synthesis with a Continuous Vocoder," in *24th European Signal Processing Conference*, Budapest, 2016.

[83] M. S. Al-Radhi, T. G. Csapó és G. Németh, „Time-domain envelope modulating the noise component of excitation in a continuous residual-based vocoder for statistical parametric speech synthesis," in *Interspeech 2017*, Stockholm, 2017.

[84] M. S. Al-Radhi, T. G. Csapó és G. Németh, „Adaptive Refinements of Pitch Tracking and HNR Estimation within a Vocoder for Statistical Parametric Speech Synthesis," *APPLIED SCIENCES,* 2019.

[85] P. Nagy, B. P. Tóth és G. Németh, „Adaptation of Large Corpus Average Voice Model in HMM Speech Synthesis for Synthesizing Short Sentences," in *Proceedings of 2nd International Acoustics and Audio Engineering Conference*, Újvidék, Szerbia, 2013.

[86] P. Nagy és G. Németh, „Improving HMM Speech Synthesis of Interrogative Sentences by Pitch Track Transformations," *Speech Communication, (82),* pp. 97-112, 2016a.

[87] G. Olaszy, G. Németh és P. Olaszi, „Automatic Prosody Generation - a Model for Hungarian," in *Eurospeech 2001*, Aalborg, Denmark, 2001a.

[88] G. Németh, G. Kiss és B. Tóth, „Cross Platform Solution of Communication and Voice/Graphical User Interface for Mobile Devices in Vehicles," in *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards*, H. Abut, J. H. L. Hansen és K. Takeda, szerk., New York, Springer, 2007, pp. 237-250.

[89] B. Tóth és G. Németh, „Challenges of Creating Multimodal Interfaces on Mobile Devices," in *Electronics in Marine International Symposium (ELMAR-2007)*, Zadar, Horvátország, 2007.

[90] B. Tóth és G. Németh, „Creating XML Based Scalable Multimodal Interfaces for Mobile Devices," in *16th IST Mobile and Wireless Communications Summit*, 2007b.

[91] S. Oviatt, A. DeAngeli és K. Kuhn, „Integration and synchronization of input modes during multimodal human-computer interaction," in *Referring Phenomena in a Multimedia Context and their Computational Treatment (ReferringPhenomena '97)*, Stroudsburg, PA, USA, 1997.

[92] S. H. Maes, „Systems and methods for synchronizing multi-modal interactions". U.S. Patent Szabadalom száma: 7,216,351, 8 May 2007.

[93] S. Schaefer, S. Bleul és W. Mueller, „Dialog Modelling for Multiple Devices and Multiple Interaction Modalities," in *Proceedings of the 2006 Workshop on Task Models & Diagrams for UI Design (TAMODIA'2006)*, Hasselt, Belgium, 2006.

[94] I. Decsi, „XML alapú multimodális felhasználói felület mobil eszközökön," BME TMIT, Budapest, 2009.

[95] G. Németh, G. Olaszy és T. G. Csapó, „Spemoticons: Text-To-Speech based emotional auditory cues," in *ICAD-2011*, Budapest, 2011.

[96] W. W. Gaver, „Auditory icons: Using sound in computer interfaces," *Human-Computer Interaction. 2,* pp. 167 - 177, 1986.

[97] S. Garzonis, J. S. J. T. és O. E. , „Auditory icon and earcon mobile service notifications: intuitiveness, learnability, memorability and preference," in *Proceedings of the 27th International Conference on Human Factors in Computing Systems*, Boston, MA, 2009.

[98] M. Jeon és B. N. Walker, „Spindex (speech index) improves auditory menu acceptance and navigation performance," *ACM Transactions on Accessible Computing (TACCESS), 3(3), 10:,* pp. 1-26, 2011.

[99] B. Tóth és G. Németh, „VoxAid 2006: Telephone Communication for Hearing and/or Vocally Impaired People," in *Computers Helping People with Special Needs*, K. Miesenberger, W. Zagler és A. Karshmer, szerk., Berlin, Springer, 2006, pp. 651-658.

[100] B. P. Tóth, P. Nagy és G. Németh, „New Features in the VoxAid Communication Aid for Speech Impaired People," in *ICCHP 2012*, Linz, 2012.

[101] G. Olaszy és G. Németh, „Voxaid: an interactive speaking comunication aid software for the speech impaired," in *Proceedings of Eurospeech '93*, Berlin, 1993.

[102] P. Mihajlik, Z. Tobler, Z. Tüske és G. G., „Evaluation and Optimization of Noise Robust Front-End Technologies for the Automatic Recognition of Hungarian Telephone Speech," in *Proc. Interspeech 2005*, Lisszabon, 2005.

[103] M. Karjalainen, P. Boda, P. Somervuo és T. Altosaar, „ Applications for the Hearing-Impaired: Evaluation of Finnish Phoneme Recognition Methods," in *Proc. Eurospeech '97 Volume 4, 1997*, Rhodes, Greece, 1997.

[104] G. Németh, C. Zainkó, B. Bogár, Z. Szendrényi, P. Olaszi és T. Ferenczi, „Elektronikus.levél felolvasó," in *Beszédkutatás '98*, Budapest, MTA Nyelvtudományi Intézet, 1998, pp. 189-203.

[105] G. Németh, C. Zainkó, G. Olaszy és G. Prószéky, „Problems of Creating a Flexible E-mail Reader for Hungarian," in *Proceedings of the 6th European Conference on Speech Communication and Technology*, Budapest, 1999.

[106] G. Kiss és G. Németh, „Gépi tanuló algoritmus automatikus címkézésre és alkalmazása beszédszintézis céljára," *Híradástechnika, LXI./3,* pp. 51-58, 2006.

[107] G. Németh és C. Zainkó, „Multilingual Statistical Text Analysis, Zipf's Law and Hungarian Speech Generation," *ACTA LINGUISTICA HUNGARICA / ACTA LINGUISTICA ACADEMICA,* pp. 385-405, 2002.

[108] H. E. Blanchard és S. H. Lewis, „Voice Messaging User Interface," in *Human Factors and Voice Interactive Systems 2nd edition*, D. G. Bonneau és H. Blanchard, szerk., New York, Springer US, 2008, pp. 257-284.

[109] P. Rutten és J. Fackrell, „The application of interactive speech unit selection in TTS systems," in *Interspeech 2003*, 2003.

[110] P. Rutten és D. Talkin, „rVoice Studio and Active Prompts," in *Speech Synthesis Workshop (SSW5)*, 2004.

[111] G. Németh, C. Zainkó, M. Bartalis és G. Olaszy, „Többnyelvű vasúti hangos utastájékoztatás korpusz alapú TTS módszerrel," *BESZÉDKUTATÁS 23,* pp. 233-241, 2015.

[112] E. Klabbers, „High-quality speech output generation through advanced phrase concatenation," in *Proceedings of the COST Workshop on Speech Technology in the Public Telephone Network: Where are we today?*, Rhodes, Greece, 1997.

[113] J. Kominek és A. W. Black, „CMU ARCTICdatabases for speech synthesis," Carnegie Mellon University, Pittsburgh, USA, 2003.

[114] G. T. B. Sárosi, A. Balog, T. Mozsolics, P. Mihajlik és T. Fegyó, „On modeling non-word events in Large Vocabulary Continuous Speech Recognition," in *2012 IEEE 3rd International Conference on Cognitive Infocommunications*, 2012.

[115] E. Csala, G. Németh és C. Zainkó, „Application of the NAO humanoid robot in the treatment of marrow-transplanted children," in *3rd IEEE International Conference on Cognitive Infocommunications*, Kassa, 2012.

[116] P. Baranyi, G. Németh és P. Korondi, „"3D Internet" alapú kognitív infokommunikáció," *HIRADÁSTECHNIKA,* pp. 70-77, 2009.

[117] Aldeberan, „NAO - Technical overview,” http://doc.aldebaran.com/2-1/family/robots/index_robots.html, 2019.

[118] P. Baxter, T. Belpaeme, L. Canamero, P. Cosi, Y. Demiris és V. Enescu, „Long-Term Human-Robot Interaction with Young Users,” in *IEEE/ACM Human-Robot Interaction 2011 Conference (Robots with Children Workshop)*, 2011.

[119] G. Olaszy, G. Németh, M. Bartalis, G. ,. Z. C. Kiss, T.-. Fegyó, G. Árvay, Z. Szepezdi és B. M. Terplánné, „Kísérleti gyógyszerinformációs rendszer beszédmodulokkal,” *Híradástechnika, LXI : 3,* pp. 8-13, 2006.

[120] G. Németh, G. Olaszy, M. Bartalis, G. Kiss, C. Zainkó és P. Mihajlik, „Speech based Drug Information System for Aged and Visually Impaired Persons,” in *Interspeech 2007*, 2007.

[121] Európai Bizottság, „Egyszerűbb hozzáférés a gyógyászati termékek adataihoz Magyarországon,” 2009.

[122] D. G. f. R. P. European Commission, „Medical products given a voice in Hungary,” in *Investing in our regions, Examples of projects co-funded by European regional policy*, Brussels, European Commission, 2010, pp. 108-109.

[123] D. EIT és W. Startups!, „https://www.eitdigital.eu/news-events/news/article/wantedeuropean-startups/,” 01 07 2015. [Online]. Available: https://www.eitdigital.eu/news-events/news/article/wantedeuropean-startups/. [Hozzáférés dátuma: 31 07 2019].

[124] A. Teixeira, A. Hämäläinen, J. Avelar, N. Almeida, G. Németh, T. Fegyó, C. Zainkó, T. Csapó, B. Tóth, A. Oliveira és e. al., „Speech-centric Multimodal Interaction for Easy-to-access Online Services A Personal Life Assistant for the Elderly,” *Procedia Computer Science,* p. 389 – 397, 2014.

[125] L. B. Christensen, „RoboBraille - Automated Braille Translation by Means of an E-Mail Robot.,” in *ICCHP*, 2006.

[126] B. Tóth és G. Németh, „Speech Enabled GPS Based Navigation System for Blind People on Symbian Based Mobile devices in Hungarian,” in *Proceedings of Regional Conference on Embedded and Ambient Systems*, Budapest, 2008b.

[127] Á. Viktóriusz, „GPS alapú navigációs rendszer vak és gyengén látó felhasználók számára Symbian alapú okostelefonokra,” BME TMIT, Budapest, 2008.

[128] B. P. Tóth és G. Németh, „Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis,” *ACTA CYBERNETICA-SZEGED,* pp. 19:(4) pp. 715-731, 2010.

[129] P. Nagy, C. Zainkó és G. Németh, „Synthesis of Speaking Styles with Corpus- and HMM-Based Approaches,” in *6th IEEE International Conference on Cognitive Infocommunications*, Győr, 2015.

[130] G. Olaszy és P. Olaszi, „Hangidőtartamok mesterséges változtatása periódusok kivágásával, megismétlésével,” in *Beszédkutatás '98*, M. Gósy, Szerk., Budapest, MTA Nyelvtudományi Intézet, 1998, pp. 151-162.

[131] P. Narasimhan, „Trinetra: "Assistive Technologies for Grocery Shopping for the Blind",” in *Tenth IEEE International Symposium on Wearable Computers*, Montreux, Switzerland, 2006.

[132] B. P. Tóth, P. Nagy és G. Németh, „Towards Modeling Interrogative Sentences in HMM-based Speech Synthesis,” *PHONETICIAN, 109-110:(I-II),* pp. 24-42, 2014.

[133] T. G. Csapó, G. Németh és M. Fék, „Szövegfelolvasó természetességének növelése,” *HIRADÁSTECHNIKA, LXIII:(5),* pp. 7-11, 2008.

A szerző tudományos közleményeinek teljes listája megtalálható az MTMT adatbázisban: https://m2.mtmt.hu/gui2/?type=authors&mode=browse&sel=10009682&view=dataSheet