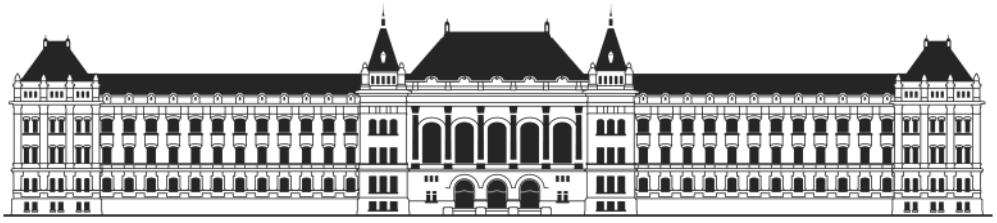
Célorientált gépi beszédkeltés interakciós rendszerekben

Az MTA doktora cím
elnyerése érdekében benyújtott értekezés tézisei

Németh Géza,
okleveles villamosmérnök, PhD



Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék

Budapest, 2019.

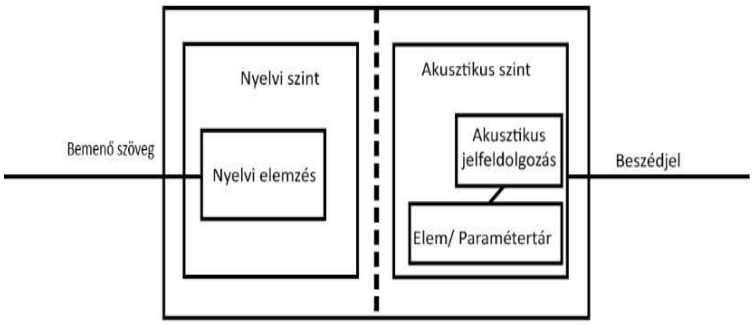
Tartalomjegyzék

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 [B1]. 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 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 [1], [2], [3]. A hangképzés artikulációs (forrás-szűrő) modellezése sikerre vezetett, hiszen a modellel az emberi beszédhez megtévesztésig hasonló hangjelenséget is sikerült létrehozni [4], 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 [5], [6]. 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) [J8]. További kutatásaink eredményeképpen szűk tématerületen (pl. időjárás jelentés, menetrend-felolvasás) létrehoztunk az emberi felolvasás minőségét és jellemzőit megközelítő

rendszereket [7]. 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 [8], [9] és részesei [10], [11], stb.

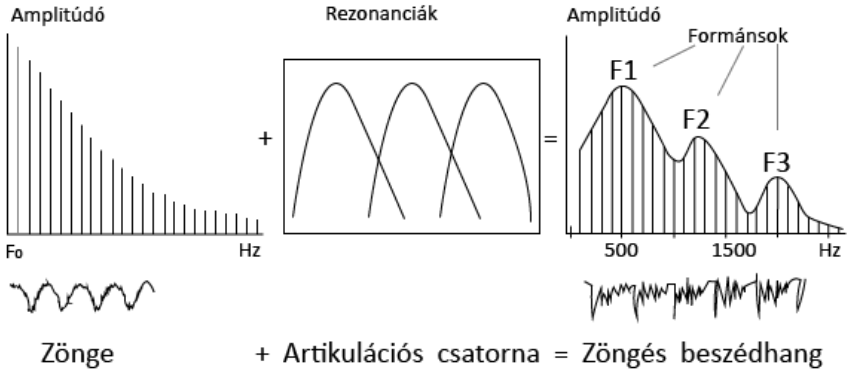
Az is kezd körvonalazódni a kutatások tapasztalatai alapján, hogy az alkalmazási területtől, az ember-gép kapcsolat megoldásától, a felhasználói elvárásoktól függően változhat a géppel előállított beszéd minőségi követelménye. Például egy beszélő robot (bábu, guruló robot) esetén az érthetőség a legfontosabb és kimondottan előnyös lehet, ha nem tökéletesen emberi jellegű, hanem robotos hangzású az előállított hang. A robotikából jól ismert a rejtélyes völgy (uncanny valley, [12]) 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 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 [13], [14], [15], [16] és [17].

A tézisfüzet 2. szakaszában történelmi áttekintés keretében ismertetem a gépi beszédkeltés különböző megközelítéseit. A 3. fejezetben kutatási célkitűzéseimet foglalom össze. A 4. szakaszban a kutatás eszközeit és módszereit tekintem át. Ezután kutatási eredményeimet foglalom össze téziscsoportonként egységes szerkezetben. Majd a korábban ismertetett tézisek gyakorlati alkalmazásokban és műszaki alkotásokban megtestesülő felhasználását tekintem át. A bevezetésnek és az azt követő következő történelmi áttekintésnek bővített változatát [18] 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. [19]. Az első elektromechanikus beszélőgép elvi módszerét is magyar ember találta fel [20]. Nagy média nyilvánosságot kapott a Bell Laboratóriumban az 1930-as években fejlesztett elektromechanikus VODER rendszer [21]. A számítógépes gépi beszédkeltés első megoldásai az 1950-es években születtek meg [22]. A mini- és mikroszámítógépek megjelenésével a hazai kutatók is követhették a nemzetközi trendeket [23], [24], [B1].

A különböző elvi megközelítések különböző beszédminőséget és gyakorlati alkalmazási lehetőségeket eredményeztek. Az artikulációs (forrás-szűrő) [25] 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 formáns-alapú beszédszintézissel sikerült kötetlen szókészletű, jól érthető, kereskedelmi forgalmazásra is alkalmas, de egyértelműen gépies hangzású, szintetizált beszédet előállítani [26]. 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 természetes beszéd hullámformájából kiinduló megoldások kutatása [5]. 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 [J8]. 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. 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 [27], [7]. Ez a terület képezi téziseim második csoportját.

A gépi beszédkeltés terén az elmúlt években – számos előnyének köszönhetően – a statisztikai parametrikus beszédszintézis vált az egyik legaktívabb kutatási területté [8]. 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.

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 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 [28], [29]. 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.

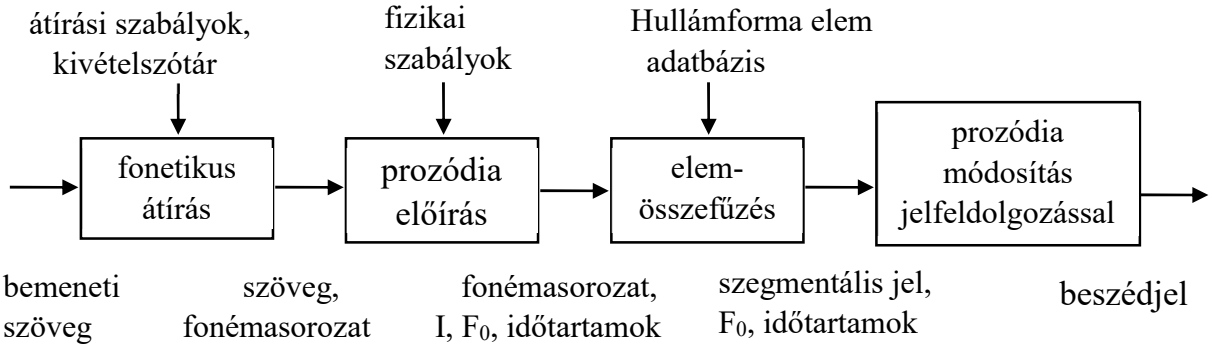
- *Multimodális információs rendszerek hatékony megoldásainak kutatása (IV. téziscsoport, alkalmazások és műszaki alkotások).*

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. 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 (ld. 4. ábra) 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 [8] az első megvalósítók között hivatkozik megoldásunkra. A III. téziscsoport eredményeit tartalmazó előadásunk [30] 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 és 1999-ben mi rendezhettük meg.



4. ábra. Hullámforma elemösszefűzésen alapuló beszédszintetizátor egy lehetséges modellje. [J8]

4. Eszközök és módszerek

A következőkben 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. Ezután következhetett a fejlesztői környezet [31] 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 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 [7]. 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).

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 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.

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), [B4].

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 [32] 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 [J8]. A kísérletekhez a GÁBOR hang diádos adatbázisát használtam fel.

4.2. A kutatások során felhasznált eszközök

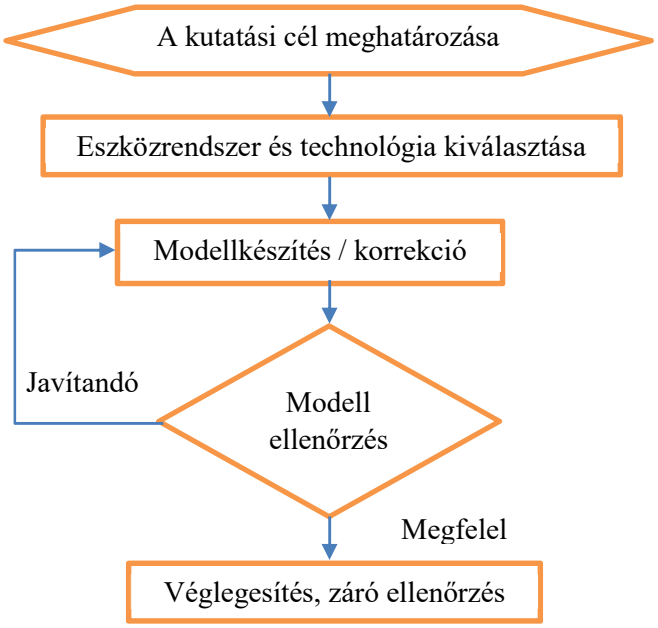
Kutatásaimhoz részben szabadon hozzáférhető eszközöket, részben pedig a BME-TMIT-en készült megoldásokat használtam. Ezek a következők:

- **VoXerver:** magyar nyelvű, automatikus beszédfelismerő beszéd-szöveg átalakítás, ill. kényszerített felismerés (Forced Alignment) üzemmódban. [33]
- **MVoxDev:** integrált szövegfelolvasó fejlesztői környezet [31]
- **Praat**: hullámforma elemzés és címkézés szoftver eszköze. [34]
- **HTS**: rejtett Markov-modell alapú gépi szövegfelolvasás keretrendszere [35]
- **DNN:** jellemző mély tanulási keretrendszerek, Merlin [36], Keras [37], Tensorflow [38], 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 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). 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 [39]) 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 tézisek összefoglalása egységes szerkezetben

I. téziscsoport: A diád és triád elemek összefűzésén alapuló gépi szövegfelolvasás

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 [5]. 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).

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. 4. 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 [40]*, Brailab* [41]*, PC talker* [42]*) Az eljárást kiterjesztettem német nyelvre is.*
Alátámasztó irodalmak: [J8], [B1]

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. Alátámasztó irodalmak: [J6], [B1]

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

A 90-es évek második felében kezdett megfogalmazódni az a koncepció, amit korpusz-alapú beszédszintézisnek nevezünk [27]. 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.

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. Alátámasztó irodalmak: [J7], [C6], [B1]

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. Alátámasztó irodalmak: [C6], [B1]

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.
Alátámasztó irodalmak: [C7], [J4], [C4]

III. téziscsoport: Statisztikus parametrikus gépi szövegfelolvasó rendszerek

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 [43]. Az IBM-nél Fred Jelinek és kutatócsoportja dolgozta ki ezen elmélet alapján az első gépi beszédfelismerő rendszert a 70-es években [44]. 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 [35].

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.

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. Alátámasztó irodalmak: [J5], [J6]

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. Alátámasztó irodalom: [C3]

III.3. Tézis: Rövid és kérdő mondatok jobb minőségű megvalósítása
Kidolgoztam a magyar kérdő mondatok alapfrekvencia-idő függvényeinek statisztikai modellezését gépi beszédelőállításhoz. Alátámasztó irodalmak: [C1], [B4]

IV. téziscsoport: Multimodális beszédinformációs rendszerek

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).

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.
Alátámasztó irodalmak: [B3], [C8], [C9]

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. Alátámasztó irodalom: [C5]

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. Alátámasztó irodalmak: [C10], [C2]

6. Az eredmények alkalmazásai, műszaki alkotások

A téziseimben bemutatott új kutatási eredmények gyakorlati alkalmazásokban és műszaki alkotásokban is felhasználásra kerültek. A disszertációban négy alapvető felhasználási területet tekintek át. A közcélú beszéd-interakciós rendszerek kizárólag beszéd modalitást felhasználó megoldások, jellemzően távközlő hálózaton vagy közlekedési utastájékoztató rendszerekben. Erre példa az elektronikus levél felolvasó rendszer távközlési szolgáltatásként (MailMondó, 1999-1), [45], [46], [47], [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 [45]. 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 [47].

Tudomásunk szerint a világon először fejlesztettünk éles szolgáltatásban megjelenő, okostelefonon futó SMS-felolvasó alkalmazást az I. és a III. téziscsoportban elért eredményekre alapozva (SMSmondó, 2003-, [49], [B2]).

A II. téziscsoport eredményeit felhasználva egy távközlési szolgáltató árlistabemondó szolgáltatása jött létre [C6]. Szintén a II. téziscsoport tette lehetővé a MÁV állomások hangos utastájékoztató rendszere [50], [30] nemzetközileg is újszerű megoldását, amit ma már több mint 100 vasúti állomást és megállóhelyet támogat.

Az egészségügyi alkalmazásokról szóló alfejezetben új, innovatív, többféle modalitást kombináló megoldásokat ismertetek. A II. és a IV. téziscsoport eredményei alapján hoztuk létre a magyarul beszélő NAO robot alkalmazását kórházi környezetben [51].

A Gyógyszervonal egy automatikus gyógyszer betegtájékoztató információs rendszer [52], [53], [54], [55], melyhez valamennyi téziscsoport eredményei hozzájárultak.

Magyarországon a projekt megvalósításakor körülbelül ötezer törzskönyvezett gyógyszer volt (ami azóta csak növekedhetett), melyek engedélyezését az Országos Gyógyszerészeti Intézet

---

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

(OGYI) végzi. Évente körülbelül 400 új gyógyszer jelenik meg és hozzávetőlegesen ugyanennyit vonnak ki a forgalomból. A projekt kitűzött célja volt, 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 fogyatékos és idős emberek számára fejlesztett, valamint az általános célú rendszereinket a terjedelmi korlátokra tekintettel csak felsorolom.

6.1. Fogyatékos és idős embereket támogató szolgáltatások

Kiemelem, hogy a VoxAid alkalmazás sztrókos betegek rehabilitációjára optimalizált változata az EIT Digital európai startup ötlet versenyén III. helyezést ért el [56].

- a PAELIFE EU AAL projekt keretében idős emberek infokommunikációs szolgáltatásainak támogatására [57]2,
- VoxAid2006 prototípus siketnéma emberek telefonálásának támogatására [C10]
- VoxAid2012 prototípus beszédsérült emberek mindennapi kommunikációjának támogatására, logopédiai és afáziás betegek rehabilitációjának támogatására [C2], [56]
- 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), [58] 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 [59] [60],
- a VUK EU AAL projekt keretében látássérült emberek beltéri navigációjának támogatására (2019).

6.2. Általános információs rendszerek

- a www.metnet.hu időjárás portál, illetve a Microsoft 2013-as fejlesztői versenyén nyertes Időjárás Mindenkinek Windows8 alkalmazás,

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

- 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).

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).

Hivatkozások

[1] 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.

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

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

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

[9] 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.

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

[11] 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.

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

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

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

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

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

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

[18] 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.

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

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

[21] 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.

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

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

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

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

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

[35] 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.

[36] 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.

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

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

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

[40] 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.

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

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

[43] 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.

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

[45] 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.

[46] 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.

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

[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 2003-as éves jelentés,* Budapest: MATÁV, 2004.

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

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.

[56] 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].

[57] 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.

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

[59] 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.

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

A tézisekhez kapcsolódó alátámasztó publikációk

Könyv ill. könyvfejezet

[B1] G. Németh és G. Olaszy, szerk., A magyar beszéd, Budapest: Akadémiai Kiadó, 2010, p. 749. (Akadémiai Nívódíj, 2011)

[B2] 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.

[B3] 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.

[B4] 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.

Folyóiratcikk

[J1] P. Nagy és G. Németh, „Improving HMM Speech Synthesis of Interrogative Sentences by Pitch Track Transformations," Speech Communication, (82), pp. 97-112, 2016a.

[J2] 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.

[J3] 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.

[J4] 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.

[J5] B. P. Tóth és G. Németh, „Hidden Markov Model Based Speech Synthesis System in Hungarian," INFOCOMMUNICATIONS JOURNAL, LXIII:(7), pp. 30-34, 2008.

[J6] 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.

[J7] A. Nagy, P. Pesti, G. Németh és T. Bőhm, „Design issues of a corpus-based speech synthesizer," . HÍRADÁSTECHNIKA, LX:(6), pp. 6-12., 2005.

[J8] 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.

Konferencia kiadvány

[C1] 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.

[C2] B. P. Tóth, P. Nagy és G. Németh, „New Features in the VoxAid Communication Aid for Speech Impaired People," in ICCHP 2012, Linz, 2012.

[C3] 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.

[C4] T. G. Csapó és G. Németh, „Prozódiai változatosság rejtett Markov-modell alapú szövegfelolvasóval," in VIII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2012.

[C5] G. Németh, G. Olaszy és T. G. Csapó, „Spemoticons: Text-To-Speech based emotional auditory cues," in ICAD-2011, Budapest, 2011.

[C6] 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.

[C7] G. Németh, M. Fék és T. Csapó, „Increasing Prosodic Variability of Text-To-Speech Synthesizers," in Interspeech 2007, 2007.

[C8] B. Tóth és G. Németh, „Challenges of Creating Multimodal Interfaces on Mobile Devices," in Electronics in Marine International Symposium (ELMAR-2007), Zadar, Horvátország, 2007.

[C9] B. Tóth és G. Németh, „Creating XML Based Scalable Multimodal Interfaces for Mobile Devices," in 16th IST Mobile and Wireless Communications Summit, 2007b.

[C10] 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.

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