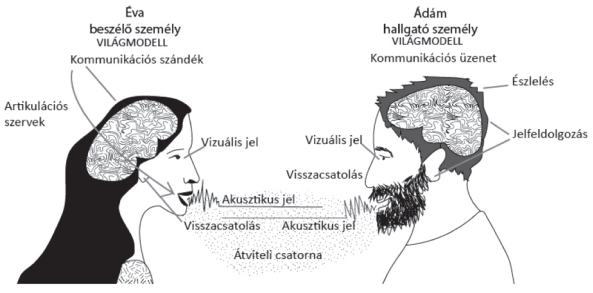
Németh Géza

Kempelentől a WaveNet-ig: a gépi beszédkeltés tudományának fejlődése

1. Bevezetés

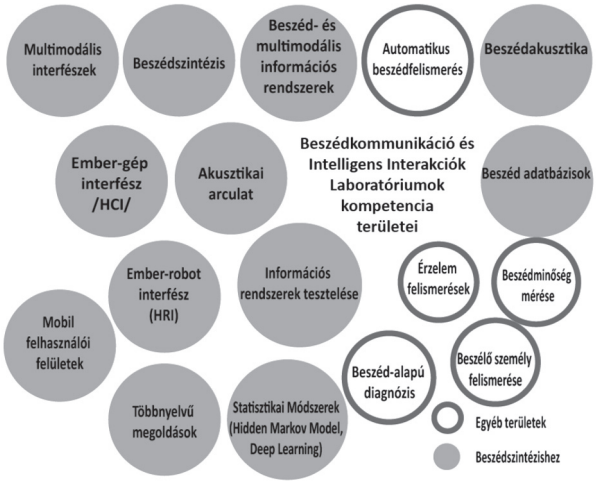
A gépi beszédkeltés a beszédtechnológia tudományterületének egyik ága. Az 1. ábrán láthatjuk a természetes beszédlánc egyszerűsített modelljét. Az emberi kommunikációnak számos alapvető feltétele van. A két partnernek a világról alkotott modellje nagymértékben meg kell egyezzen. Ez a modell hosszú időszak 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 hang- vagy videotelefonon) keresztül jutnak el a hallgatóhoz. A hallgató személy érzékszervei adják tovább a megfelelő jelfeldolgozás után az észlelés számára az információt. A kommunikációs üzenet értelmezése a hallgató személy világról



1. ábra. A természetes beszédlánc egyszerűsített modellje

alkotott modelljéhez kapcsolódóan alakul ki. A beszédkommunikáció alapvető jellemzője, hogy a beszélő és a hallgató szerep 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 (például hangerő meghatározásban). A továbbiakban az akusztikus csatorna szerepével foglalkozunk, mert a gépi feldolgozásban általában annak van elsődleges szerepe.

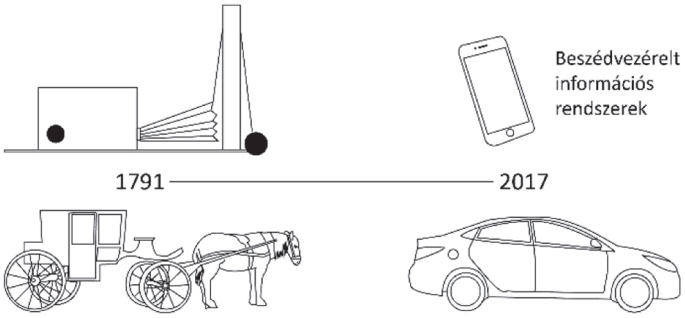
Beszédtechnológiának a természetes beszédlánc egy vagy több elemének gépi megvalósítását tekintjük.[1] A beszédtechnológia interdiszciplináris tudomány, számos bölcsészeti (például nyelvtudomány, fonetika, pszichológia), természettudományi (például fizika, matematika) és műszaki területet (például akusztika, jelfeldolgozás) érint. Laborcsoportunk kompetencia területeit mutatja a 2. ábra.



2. ábra. Laborcsoportunk kompetenciaterületei

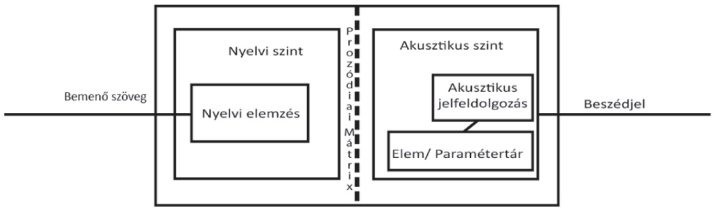
[1] Németh és Olaszy 2010.

A jelen tanulmányban a gépi beszédkeltés tudományának fejlődését tekintjük át. Ezt illusztrálja a 3. ábrán látható idővonal.



3. ábra. A közlekedés eszközeinek és a gépi beszédkeltés technológiáinak fejlődési idővonala

A gépi szövegfelolvasás általánosított modellje a 4. ábrán látható. A nyelvi szinten a bemenetre kerülő szövegből meghatározzuk a kimondandó hangokat és azok alapvető prozódiai jellemzőit (időtartam, intenzitás, zöngés hangokra alapfrekvencia-menet). Az akusztikus szinten pedig a rendelkezésre álló technológiától függő elemtár és jelfeldolgozási módszer segítségével a prozódiai mátrix adatai alapján előállítjuk a kimeneti beszédjelet.



4. ábra. A gépi szövegfelolvasás általánosított modellje

A beszédkeltés gépi modellezése több mint két évszázadra tekinthet vissza.[2] Hosszú ideig csak az artikulációs csatorna modelljének vezérlését oldották meg. 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.[3] A hangképzés artikulációs modellezése sikerre vezetett, hiszen a modellel az emberi beszédhez megtévesztésig hasonló hangjelenséget is sikerült létrehozni[4] a vezérlő paraméterek hosszadalmas kézi optimalizálása révén. A szövegfelolvasáshoz szükség volt a szöveg valós idejű beadására és számítógépes elemzésére is.[5] Azonban ezzel a megoldással a fő célt, az automatizált gépi szövegfelolvasás emberre emlékeztető szintjét nem sikerült elérni. A géppel keltett beszédjel érthető, de meglehetősen robotos hangzású volt.

Ezért az 1990-es évek elejétől előtérbe kerültek az emberi hangképzés eredményeként előálló hullámforma tárolásán, feldolgozásán, módosításán és visszajátszásán alapuló megoldások.[6] Ezek segítségével lehetett hosszabb ideig folyamatosan használható gépi felolvasó rendszereket létrehozni (például e-levél felolvasása és képernyő felolvasása látássérült emberek számára).[7] Szűk tématerületen (például időjárás jelentés, menetrend-felolvasás) kutatásaink eredményeképpen már magyar nyelven is lehet az emberi felolvasás minőségét és jellemzőit megközelítő rendszereket létrehozni.[8] Az elmúlt évtizedben pedig az artikulációs és a hullámforma-alapú megközelítés előnyeinek kombinációját ígérő statisztikai parametrikus beszédszintézis (elsősorban Hidden Markov-Model, HMM és Deep Neural Networks, DNN) kialakulásának lehettünk tanúi.[9] A legújabb technológia

[2] Kempelen 1989.
[3] Stevens, Kasowski és Fant 1953.
[4] Rosenberg, Schafer és Rabiner 1971.
[5] Olaszy 1989, Klatt és Klatt 1990.
[6] Moulines és Charpentier 1990, Beutnagel és mtsai 1999.
[7] Olaszy, Németh és Olaszi és mtsai 2000.
[8] Németh, Olaszy és Fék 2006.
[9] Zen, Tokuda és Black 2009, Zen, Senior és Schuster 2013.

(WaveNet)[10] pedig a hullámformából tanulja meg a modell paramétereit, és a megelőző néhány ezer hullámforma minta alapján ad becslést a következő mintára.

Időközben az is kezd körvonalazódni a kutatásokban, hogy az alkalmazási területtől, az ember–gép kapcsolat megoldásától függően változhat a géppel előállított beszéd minőségi követelménye. Például egy beszélő 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)[11] 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 vált ki 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 optimális ember–gép interfészt megvalósítani.

3. Áttekintés

A gépi beszéd-előállítás tudományos alapjait Kempelen Farkas 1791-ben megjelent könyve fektette le.[12] Az első elektromechanikus beszélőgépet is magyar ember találta fel.[13] Nagy médianyilvánosságot kapott a Bell Laboratóriumban az 1930-as években fejlesztett elektromechanikus VODER-rendszer.[14] A (nagy)szá-

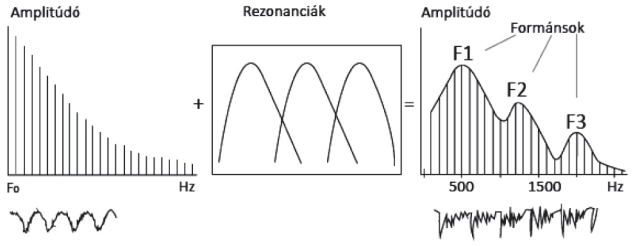
[10] Zainkó, Tóth és Németh 2017.
[11] Mori 1970/2012.
[12] Kempelen 1989.
[13] Bánó 1916.
[14] Dudley, Riesz és Watkins 1939.

mítógépes gépi beszédkeltés első megoldásai az 1950-es években születtek.[15] A mini- és mikroszámítógépek megjelenésével a hazai kutatók is követhették a nemzetközi trendeket.[16]

Az elmúlt évtizedekben a számítástechnika technológiai fejlődése a gépi szövegfelolvasás területén is több technológiai megközelítés kutatását és alkalmazását tette lehetővé. Látható, hogy a gép beszéd-előállítás témakörében a technológia tükrében mindig változó kutatási kérdések merülnek fel. Ezek megoldása folyamatos kihívást jelent, és egyrészt egymást követő alternatív tudományos generációkat eredményez. Másrészt azzal jellemezhető, hogy a korábbi generációk nem avulnak el (mind a mai napig használatban vannak), hanem az újabb generációk más-más peremfeltételek optimalizálását igénylik és teszik lehetővé.

Formánsszintézis

A különböző elvi megközelítések különböző beszédminőséget és gyakorlati alkalmazási lehetőségeket eredményeznek. Az artikulációs[17] megközelítés elsősorban az emberi beszédkeltés mechanizmusainak modellezésére volt alkalmas. A formánsalapú beszédszintézissel (lásd 5. ábra) sikerült kötetlen szókészletű, jól

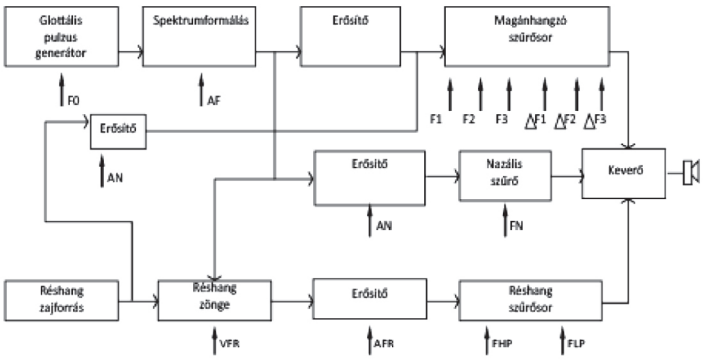


5. ábra. A gépi beszédkeltés formáns modelljének alapelve

[15] Cooper 1961.
[16] Olaszy 1978, Gordos és Takács 1983.
[17] Mermelstein 1973.

érthető, kereskedelmi forgalmazásra alkalmas, de egyértelműen gépies hangzású, gépi hangot előállítani.

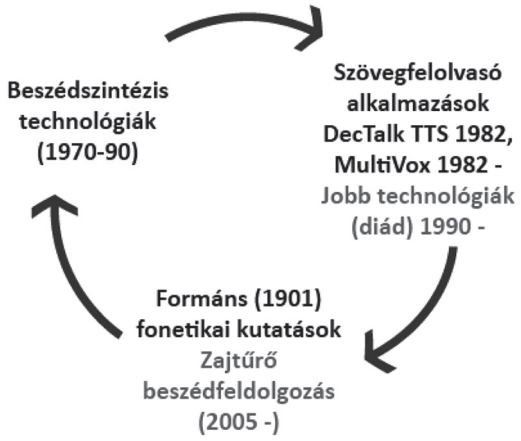
A modell lényege az ún. forrásszűrő megközelítés. A zöngés hangokat azonos alapfrekvenciájú (F0) periodikus gerjesztéssel, a zöngétleneket fehérzaj-szerű forrásjellel és az artikulációs csatornát szimuláló szűrősorral modellezzük. Az így kapott kimeneti jel hullámformája és frekvencia spektruma (azokban a formáns értékek, melyek meghatározóak a magánhangzók észlelésében) jó közelítéssel megegyezik a természetes beszédével. A 6. ábrán egy formáns modell részletes blokkdiagrammját láthatjuk.



6. ábra. Formánsszintetizátor blokkdiagramja

A mai napon (2018. március 14.) bekövetkezett haláláig ilyen elveken alapuló rendszert használt Stephen Hawking, az ismert elméleti fizikus, mert az évtizedek alatt azonosult a gép hangkarakterével.

Érdemes röviden áttekinteni azt a ciklikusságot (7. ábra), ami az elméleti kutatások alapján elért technológiai eredmények után a gyakorlati alkalmazásokig jut, majd az itt felmerülő problémák újabb elméleti megalapozottságot igényelnek.



7. ábra. A formáns témakör ciklikus fejlődése

A formánsokkal kapcsolatos elméleti kutatások a 19. század végétől erősödtek fel,[18] és az 1970-es évektől eredményeztek gyakorlati alkalmazást is ígérő beszédszintézis technológiákat.[19] Talán a legismertebb formáns alapokon nyugvó angol nyelvű szintetizátor a DecTalk volt, ami az MIT professzora, Dennis Klatt kutatásain alapult.[20] Stephen Hawking is ennek egy változatát használta.

Érdemes megjegyezni, hogy egy ilyen önálló dobozban található eszköz ára 1984-es megjelenésekor mintegy 4000 USD volt. Ezzel gyakorlatilag egy időben készült el a HungaroVox rendszer az MTA Nyelvtudományi Intézetében[21] (lásd 8. ábra).

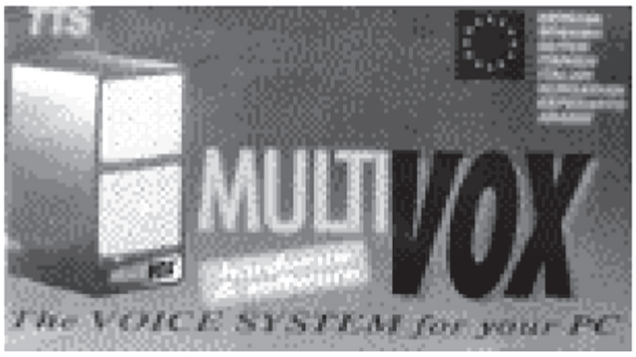
Ennek továbbfejlesztéseként készültek el a MultiVox különböző változatai[22] a BME-n előbb PC-hez illesztett hardver, illet-

[18] Lloyd 1890.
[19] Flanagan és Rabiner (eds.) 1973.
[20] D. Klatt 1987.
[21] Kiss és Olaszy 1984.
[22] Olaszy 1989.

8. ábra. A HungaroVox rendszer doboza

ve önálló dobozos eszközként, majd szabadon letölthető tisztán szoftver változatban.

A formánsokat nemcsak a beszéd szintézisére, hanem felismerésére is alkalmazták.[23] A formánsok automatikus meghatározása és transzformációja is a folyamatosan vizsgált elvi problémák közé tartozik.[24] Újabban sokféle szintetizált beszédhang előállításában és a zajos beszéd felismerésénél erősödött fel az érdeklődés ebben az irányban.

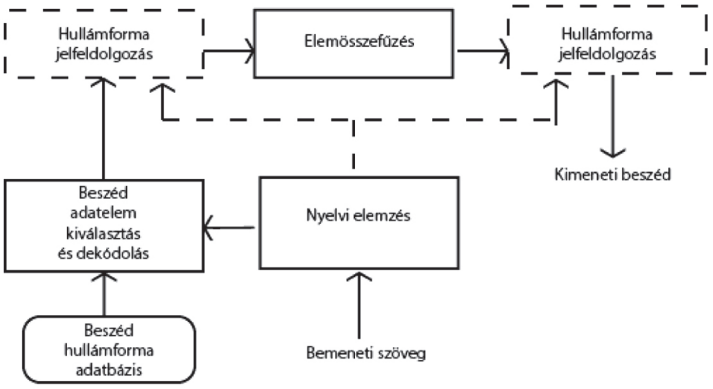


9. ábra. A MultiVox rendszer doboza

[23] Furui 2005.
[24] Bőhm és Németh 2006.

Elemösszefűzéses és elemkiválasztásos gépi szövegfelolvasás

A korábbi parametrikus (elsősorban formánsalapú) érthető, de erősen robotos hangzású szövegfelolvasási technológia továbbfejlesztésére a 80-as évek végétől alakult ki az a koncepció, hogy próbálkozzunk természetes beszéd rögzítésével, címkézésével és a visszajátszáskor a megfelelően kiválasztott elemek összefűzésével, és (ha szükséges) jelfeldolgozás segítségével történő optimalizálásával. Ennek egyik lehetséges megoldását láthatjuk a 10. ábrán. Ennek a megoldásnak az egyik formája, hogy alapelemnek hangpárokat reprezentáló beszédhullámforma-részleteket (ún. diádok, angolul diphone) választunk. Ekkor például az alma szót _a, al, lm, ma, a_ (_ a szünet jele) diádokból lehet előállítani. A magyar nyelv 39 hanggal (25 mássalhangzó –C- és 14 magánhangzó –V-) plusz a szünet (_ jel) lefedhető, tehát az adatbázisban mintegy $40^2 = 1600$ elemre van szükség. A hosszú mássalhangzókat időtartam módosítással tudjuk megoldani. A minőséget egyrészt a sok, folytonossági hibát okozó vágási pont, másrészt a prozódiát megvalósító modell egyszerűsége és a jelfel-



10. ábra. Elemösszefűzéses és elemkiválasztásos gépi szövegfelolvasó rendszer általános blokkdiagramja (a – blokkok opcionálisak)

dolgozás korlátozza.[25] A minőséget tovább javíthatja a hangok vágás nélküli összefűzését legalább a magánhangzók esetében figyelembe vevő hanghármasok (triád, angolul triphone) alkalmazása. Például ekkor az alma szót az _al, lm, ma_ két triádból és egy diádból lehet előállítani.

Ezzel a megoldással egyrészt az eredeti emberi hangszínezetre emlékeztető gépi beszédet lehet létrehozni, másrészt viszonylag kis számítási kapacitás mellett lehet változtatható hangkaraktereket (változó alapfrekvencia és beszédtempó) kialakítani. Ennek különös fontossága van a látássérült emberek kommunikációjának szempontjából.

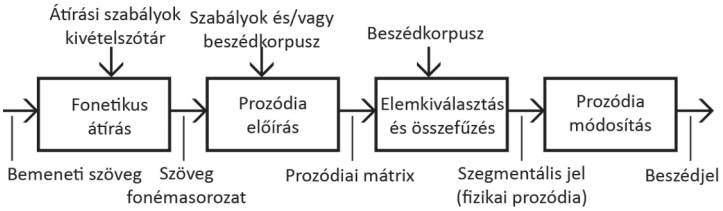
Az elemösszefűzéses megoldás hátránya, hogy egy-egy hangkapcsolathoz csak egyetlen emberi bemondásból származó mintát tárol. Ezért a beszéd megfelelő prozódiai jellemzőit az összefűzött elemeken végzett jelfeldolgozással kell biztosítani.

Elemkiválasztásos gépi szövegfelolvasás

Az első magyar nyelvű korpuszalapú, hullámforma elemválogatásra épülő gépi szövegfelolvasó rendszer modelljét láthatjuk a 11. ábrán. Szubjektív értékelés szerint az emberi felolvasáshoz jobban hasonlító hangot állít elő, mint a korábbi magyar nyelvű szövegfelolvasó rendszerek. Három célorientált alkalmazási területen (időjárás-jelentés, árlista és pályaudvari tájékoztató felolvasása) kiderült, hogy ennek a technológiának a felhasználásával lehetséges az emberi felolvasáshoz megtévesztésig hasonló magyar nyelvű gépi felolvasást létrehozni.

A 90-es évek második felében kezdett megfogalmazódni az a koncepció, amit korpuszalapú beszédszintézisnek nevezünk.[26] Az elképzelés alapötletét az az általánosan elfogadott elv adja, hogy egy hullámforma-összefűzésen alapuló szövegfelolvasó rendszer minőségét döntően az összefűzések száma és az összefűzött ele-

[25] Olaszy, Németh és Olaszi és mtsai 2000.
[26] Möbius 2000.



11. ábra. Korpusz alapú, hullámforma elemkiválasztásos beszédszintetizátor modellje (Fék és mtsai 2006)

meknek az emberi kiejtéshez való hasonlósága határozza meg. Minél hosszabb elemekből állítjuk elő a szintetizált beszédet – az összefűzési pontok számának csökkenése és a természetes beszédhez való jobb illeszkedés miatt –, annál jobb lesz az elért minőség. 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 ennél rövidebb egységeket vesznek fel az adatbázisba, de azzal a céllal, hogy nagy valószínűséggel hosszú elemekből összefűzhető legyen a kimenet.

Ennek egyik szélsőséges megoldása például az autóbuszokon alkalmazott bemondások digitális rögzítése, majd megfelelő egyszerű vezérlés (például nyomógombok) segítségével történő visszajátszása. Például: *A következő megálló – a Keleti pályaudvar*. A mondat első fele a hangos bemondásban 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). Természetesen ennek a megoldásnak egyrészt jelentős a tárigénye, másrészt erősen korlátozott a témaköre.

A fenti koncepció alapján külföldön már készült néhány korpuszalapú beszédszintetizátor a világnyelvekre,[27] magyar megol-

[27] Möbius 2000, Kawai és mtsai 2004.

dást azonban elsőként a BME-n hoztunk létre. Munkánk során felhasználtuk a korábbi magyar nyelvű kutatások[28] 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 szóalapú megközelítések nem alkalmazhatók közvetlenül.

Első kísérleti területünk az időjárás jelentés témaköre volt. Húsz internetes oldal 2004 áprilisa és 2005 májusa közötti időjárás-jelentéseinek alapján reprezentatív szöveges adatbázis jött létre (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 a következő modell vált be: 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 későbbiekben részletezett prozódiai változatosság is megoldott legyen.

A szintézis optimális alapelemének a szóelemet választottuk (két szóköz közötti karaktersorozat), valamint az ebből felépülő hosszabb szövegrészeket (szófüzér, önálló mondatrész stb.). A szó méretű elem egyrészről hosszabb a diád-triád elemeknél, tehát akusztikai tartalma biztosan jobban képviseli az optimális hullámformát, másrészről a percepciós feldolgozásunk során az anyanyelvi bázisunk a szó feldolgozására épül az agyunkban. Ha tehát jó akusztikai tartalmú szó kerül a szintetizálandó mondatba, akkor természetesebb hangzásúnak fogjuk ítélni, mint a diád/triádokból összerakott ugyanazon szót. Mindezt segíti, ha prozódiai szempontból is megfelelő szó kerül a szintetizálandó mondat adott helyére. Mindezekből adódik, hogy a szintézishez

[28] Olaszy és Németh 1999.

használt beszédadatbázisnak két kritériumnak kell eleget tennie. Az első az, hogy minden szóból legalább háromfélét kell tartalmaznia (mondatkezdő, -belső és -záró elem). A második az, hogy tartalmazzon megfelelő diád/triád lefedettséget is tetszőleges szöveg (az optimálisnál rosszabb, de lehetséges) előállításához

A prozódia modellben az alapegység a mondat. A modell szorosan összefügg a szintetizálandó szöveg szerkezetével. Jelen esetben döntően kijelentő mondatokat modellezünk. A kijelentő mondat prozódiai szerkezete jól körülhatárolható, ismert egységekből áll. Ezeket az egységeket a mondaton belüli hely szerinti pozicionálással (hol van a szó a mondatban), valamint a központozással (vesszők, gondolatjelek stb.) kapcsolatba lehet hozni. Ez a modell lényege. Ugyanaz a modell kerül alkalmazásra a szöveges adatbázisban, a beszédadatbázisban és a szintetizálandó mondatban is. Alkalmazásával nincs szükség prozódiai jellegű jelfeldolgozás használatára a szintézis során Az ezeknek a peremfeltételeknek megfelelően mohó algoritmussal[29] kialakított szöveges adatbázis rugalmasan bővíthető. Végül 5821 mondatot, 102 940 szót tartalmaz, ami 488 093 hangnak (fonémának) felel meg.

A szöveges adatbázist mintegy két hónapos munkával egy professzionális bemondó felolvasásában rögzítettük. Ezután utófeldolgozás következett. A hullámformát több szinten címkéztük. A legalsó szinten fonéma (hang) címkékkel történő ellátáshoz félautomatikus eljárás valósult meg a BME TMIT-en fejlesztett beszédfelismerő[30] felhasználásával. A beszédfelismerőt ún. kényszerített üzemmódban (forced alignment, az ismert szövegnek megfelelő hangok pozícióját kellett megjelölni a hullámformában) használtuk.

A prozódiai modulban a magyar nyelvhez és a célorientált megközelítéshez illeszkedő új indirekt eljárást valósult meg. Az adott szó mondatban, illetve prozódiai egységben elfoglalt helyéhez lehetőleg optimálisan illeszkedő elemek (elsősorban szó,

[29] Cormen, Leiserson és Rivest 1990.
[30] Mihajlik és mtsai 2007.

ha az nincs, diád/triád hangelemek) összefűzését végzi az algoritmus.

Az elemkiválasztás és összefűzés modulban két költségfüggvény összegének minimalizálása valósul meg új, fonetikai szempontok szerint kialakított költségfüggvények alapján. Az egyik költségfüggvény az egyes elemek (szó- és hangszinten eltérő) egymáshoz illeszkedésének (folytonosságának) felel meg (ún. összefűzési költség). Mivel a kiejtés folyamatos, a (szó)határokon törekedni kell arra, hogy a spektrális illeszkedés (például formánsmenet) is folyamatos legyen. A szavak első és utolsó hangjának illeszkedése kerül vizsgálatra, és az illeszkedés költsége több szempont alapján számítódik 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 beszédkorpuszban, 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 beszédkorpusz elég nagy) ez meg is valósul, így a szintetizált szöveg hangzása közel lesz a természeteshez.

A másik költségfüggvény határozza meg, hogy hangsor és hangkörnyezet szempontjából a kiválasztott elem (szó, szófüzér vagy hang) mennyire felel meg a prozódiai követelményeknek. 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 prozódiai költség meghatározásánál – az időtengelyi pozíción felül – felhasználjuk az alapfrekvencia (F_0) értékének a változását is. Ha nagy F_0 ugrás van a két elem között, akkor a költség magas lesz, tehát a két elem nem illeszthető össze.

A költségfüggvények súly értékeit iteratív módon, mintegy 500 mondat többszöri szintézisével határoztuk meg. A költségfüggvények alapján először a szószintű, majd a hangszintű optimális elemeket választjuk ki Viterbi-algoritmus segítségével. 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.

Ideális esetben prozódiai módosítást végző jelfeldolgozás nélkül történik az összefűzés.[31] Ebben a megoldásban egy-egy hangkapcsolathoz akár több tízezer minta is tartozhat. A nehézsége a megnövekedett tárhelyigényen túlmenően az, hogy a hatalmas számú alternatíva közül kell valós időben megtalálni a (közel) optimális megoldást.

A hullámforma alapú megoldások hátránya, hogy a legjobb minőséghez minden egyes beszélőtől külön-külön beszédadatbázist kell rögzíteni. Beszédadatbázison a következőt értjük: hanganyag, azaz a felolvasásból származó emberi beszéd, az elhangzott szöveg fonetikus átirata és többszintű szegmentálási címkék halmaza. A beszédadatbázist (más néven beszédkorpuszt) jellemzően az adott kutatási feladathoz illesztve készítik el. Ez nemcsak hangfelvételt jelent, hanem az adott technológiától függő összetett címkézési feladatot is. Ha a beszédtempót változtatni akarjuk, az csak jelfeldolgozási műveletekkel lehetséges, amik rontják a beszédjel minőségét. Ennek a technológiának a legújabb hazai felhasználási területe a MÁV-állomásokon hallható magyar és angol nyelvű célorientált gépi szövegfelolvasó rendszer.[32] A mai napig ez a technológia szolgáltatja a legjobb beszédminőséget.

Statisztikus parametrikus gépi felolvasó rendszerek

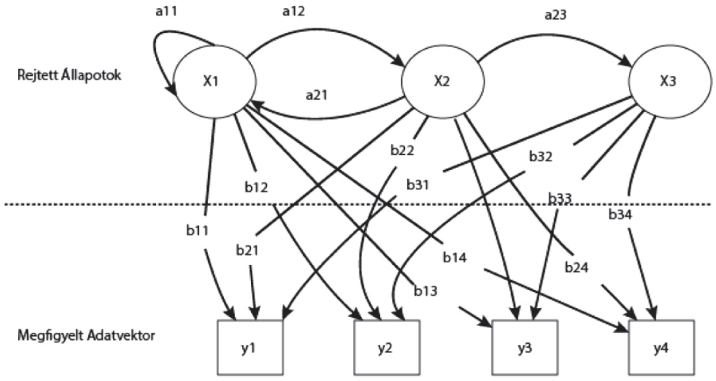
A gépi beszédkeltés terén az elmúlt években – számos előnyének köszönhetően – a statisztikai parametrikus beszédszintézis vált az egyik legaktívabb kutatási területté.[33] Ennek során először kinyerjük a jellemző paramétereket (például spektrális összetevők,

[31] Möbius 2000, Németh, Olaszy és Fék 2006.
[32] Zainkó, Bartalis és mtsai 2015.
[33] Zen, Tokuda és Black 2009.

alapfrekvencia, hangidőtartamok, hangok elhelyezkedése, hangkörnyezet) a beszédkorpuszból, majd ezen paraméterek sokaságát HMM és DNN modellekkel helyettesítjük.

Rejtett Markov-modell alapú gépi beszédkeltés

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. A 12. ábra szemlélteti a HMM-modell alapgondolatát.

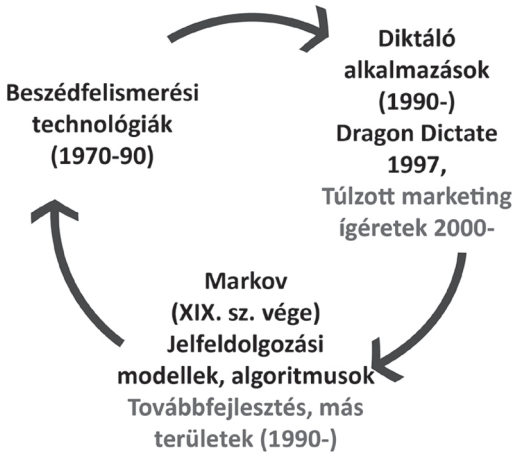


12. ábra. A HMM-modell alapgondolata

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 állapot átmenetek 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 (például cepstrális együtthatók) az adatvektorok, az állapotok pedig a kimondott beszédhangoknak felelnek meg. Beszédszintézis során pedig az adatvektorokat a bemeneti szöveghez tartozó intonációs mátrix adatai jelentik, az állapotoknak pedig egy beszédkódoló paramétervektorai (egyebek között spektrális adatok) felelnek meg.

A HMM témakörben is megfigyelhető az alapkutatás – technológia kutatás-fejlesztés – alkalmazások ciklikussága (13. ábra). Az alapelveket Andrej Markov, orosz matematikus dolgozta ki a 19. század végén és a 20. század elején.[34] Az 1960-as években merült fel az elmélet gyakorlati felhasználása.[35] Az IBM-nél Fred Jelinek és kutatócsoportja dolgozta ki

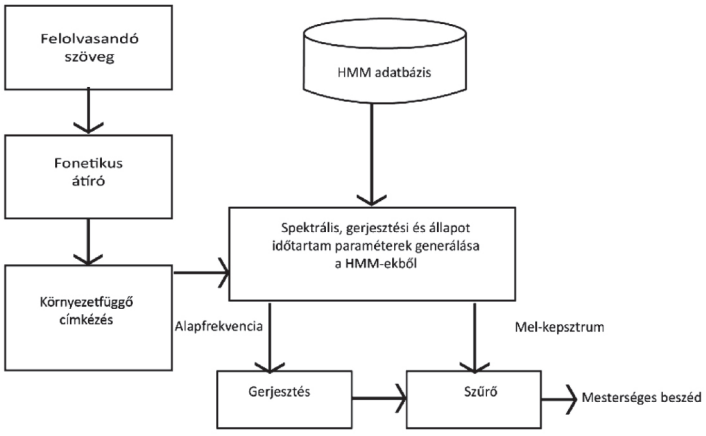


13. ábra. A HMM-kutatások ciklusai

[34] Markov 1913.
[35] Baum és Petrie 1966.

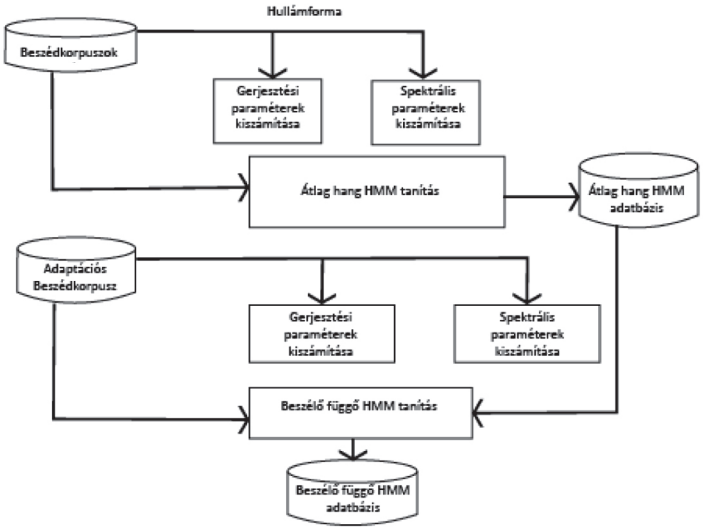
ezen elmélet alapján az első gépi beszédfelismerő rendszert a 70-es években.[36] Ennek alapján jöttek létre az első kereskedelemben kapható nagyszótárú beszédfelismerő rendszerek (IBM Tangora, Dragon Systems, Philips dictation stb.). A beszédfelismerésben elért sikerek vezettek oda, hogy felmerült az elmélet alkalmazása gépi szövegfelolvasás céljaira is. Az első ilyen rendszert a nagoyai egyetemen Tokuda professzor irányításával fejlesztették ki.[37] Természetesen ez jelentős további alapkutatási feladatokat vetett fel.

Egy HMM-alapú gépi szövegfelolvasó rendszer blokkdiagramját látjuk a 14. ábrán.[38] A rendszer előnye, hogy megfelelően felcímkézett adatbázisból automatikusan állítható elő a HMM-modellek adatbázisa. Ez jelentős számítási időt igényel. Viszont a modellek már gyorsan generálják a beszédkódoló vezérléséhez szükséges paramétersorozatot. Ez lényegesen hatékonyabb, mint



14. ábra. A HMM-szintézis alapelve

[36] Jelinek 1976.
[37] Tokuda és mtsai 2000.
[38] Tóth és Németh 2008.



15. ábra. A HMM-hangadaptáció blokkdiagramja

a hullámforma elem összefűzéses vagy elemkiválasztásos megoldások. Így megnyílt az út a sok hangon beszélni képes gépi szövegfelolvasó rendszerek kialakítása előtt.

További előnyöket jelent az, hogy viszonylag rövid (akár néhány percnyi) hangminta alapján is lehetséges az adott beszélő hangjára emlékeztető felolvasó rendszer létrehozása.[39] Ehhez célszerűen mintegy tíz beszélő személy nagyobb (személyenként kb. 2000 mondatot tartalmazó) hangadatbázisából egy ún. átlag hang-HMM-modellt hozunk létre az adott nyelven. Majd ezt a modellt tudjuk a rövid felvétel alapján az adott beszélőhöz igazítani. Ennek a megoldásnak a blokkdiagramja látható a 15. ábrán.

[39] Tóth és Németh 2010.

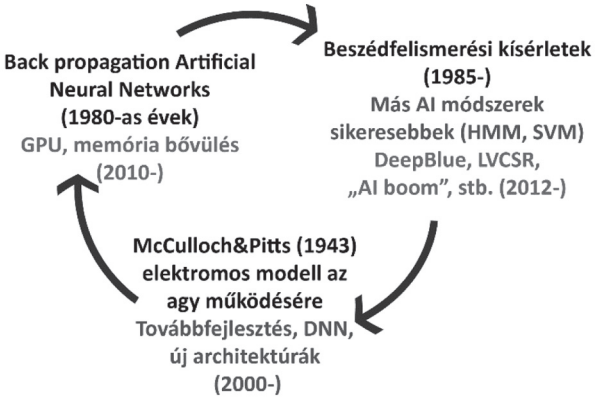
Neurális hálózatok

A neurális hálózatok elmélete is meglehetősen hosszú időre tekinthet vissza.[40] A gyakorlati alkalmazáshoz nagy lökést adott a modellek tanítására kidolgozott eljárás.[41] A számítástechnika gyors fejlődése is népszerűsítette ezt a megközelítést a 80-as években. A beszédfelismerés területén a HMM versenytársaként tekintettek erre az alternatívára.[42] Azonban az adott technikai korlátok között a HMM jobb eredményeket adott. Ezért erősen lecsökkent a téma iránti lelkesedés, de az elméleti munka tovább folytatódott. Többek között ez is vezetett az IBM Deep-Blue rendszere kifejlesztéséhez, ami 1997-ben legyőzte az aktuális sakkvilágbajnokot.[43] A számítási kapacitás növekedése (különösen a grafikus jelfeldolgozó kártyák – GPU – megjelenése) és a memóriaméretek megsokszorozódása elhozta a sok réteget tartalmazó mély neurális hálók (deep neural networks, DNN) felhasználási lehetőségét is a nagyszótárú gépi beszédfelismerés területén a HMM-nél jobb eredményekkel.[44]

Ezután természetszerűleg kezdték el alkalmazni a DNN-megoldásokat a gépi szövegfelolvasás területén is, például a spektrális és a prozódiai paraméterek becsléséhez.[45] Ezt a folyamatot illusztrálja a 16. ábra.

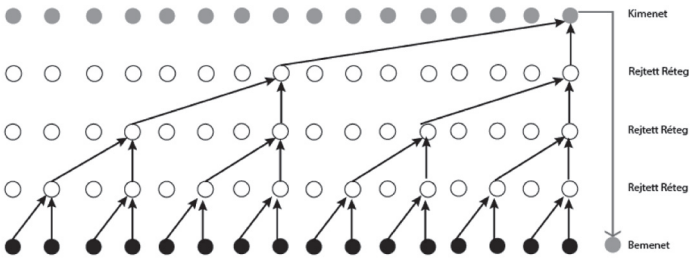
A memória és GPU-kapacitás rohamos növekedése és a felhőalapú infrastruktúrákban összpontosuló hatalmas erőforrások vezettek korábban elképzelhetetlen számítás igényű modell tanítási módszerek kidolgozásához. 2016 januárjában jelent meg az első publikáció arról, hogy pixelenként tanítanak be pixelbecslő neurális hálózatot.[46] Ugyanezen év szeptemberében ez az elv már

[40] McCulloch és Pitts 1943.
[41] Werbos 1974.
[42] Waibel és mtsai 1989.
[43] Campbell, Hoane és Hsu 2002.
[44] Dahl és mtsai 2012.
[45] Zen, Senior és Schuster 2013, Nagy és Németh 2016.
[46] Van den Oord, Kalchbrenner és Kavukcuoglu 2016.



16. ábra. A neurális hálózati kutatások ciklusai

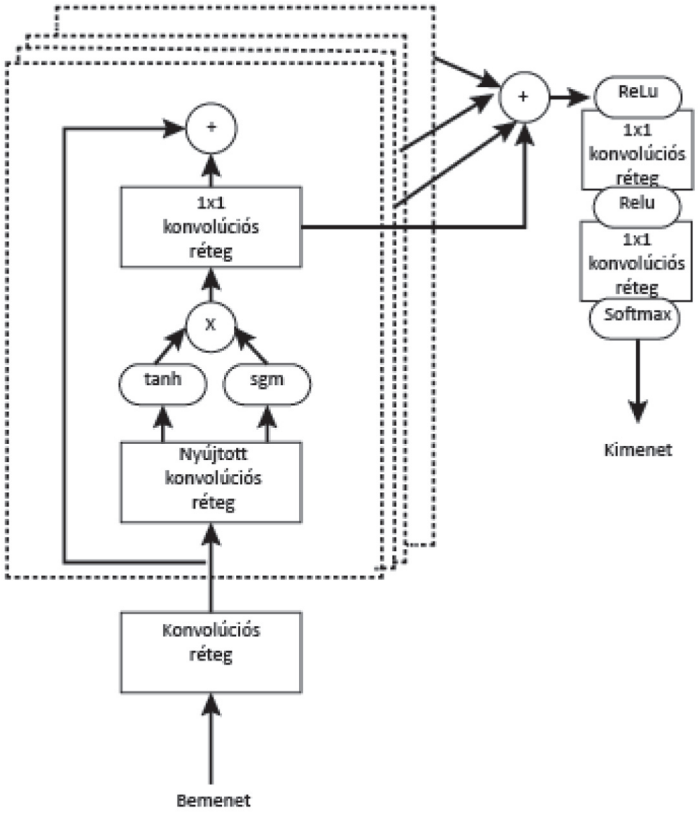
a gépi szövegfelolvasás témakörében került alkalmazásra.[47] Nem lényegkiemelt paramétereket, hanem a következő mintát becsli néhány ezer előző minta alapján a hálózat (lásd 17. ábra).



17. ábra. A WaveNet alapelve
(az ábra forrása: https://deepmind.com/blog/wavenet-generative-model-raw-audio/)

A beszéd mellett zene generálására is alkalmazható. Büszkék vagyunk arra, hogy kutatócsoportunk néhány hónap alatt magyar

[47] Van den Oord és mtsai 2016.



18. ábra. Magyar WaveNet TTS blokkdiagram
(Zainkó, Tóth és Németh 2017)

nyelvre is alkalmazta ezt a megközelítést[48] a 18. ábrán látható felépítésben.

A terület rohamos fejlődését jelzi, hogy a Google személyi asszisztensében 2017 szeptemberében (az első cikk megjelenése után nagyjából egy évvel) az amerikai angol és a japán nyelvű termék ezen megközelítés alapján generált hangon szólalt meg.

[48] Zainkó, Tóth és Németh 2017.

3. Meghallgatásos tesztek, értékelések

A gépi szövegfelolvasás és a felhasználói felületek értékelésében általánosan elterjedt az eredmények MOS (Mean Opinion Score) és CMOS (Comparison Mean Opinion Score) alapú értékelése. MOS-alapú teszt esetén a tesztalanyok a mintákat 1-től (legrosszabb) 5-ig (legjobb) értékelhetik (egész számokkal), CMOS esetén pedig szintén ötelemű 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 bízzuk. 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 tesztalanyban keltett érzelem stb. szerepet játszik.

A tanulmányban ismertetett technológiák hangmintái meghallgathatók a http://smartlab.tmit.bme.hu és a http://magyarbeszed.tmit.bme.hu honlapokon.

4. Az eredmények alkalmazhatósága

Az áttekintett gépi beszédkeltési technológiák ilyen vagy olyan szempontból ma is használhatók, de újabb elméleti és gyakorlati problémákat vetnek fel. Tehát szó sincs arról, hogy a beszédszintézis témaköre megoldottnak lenne tekinthető. Néhány megoldásra váró kutatási probléma: az emberi beszédhez hasonló változatosság (minden emberi megszólalás egyedi és megismételhetetlen), az adott kommunikációs kontextushoz illő beszédstílus alkalmazása, gyors adaptáció új témakörökhöz.

Viszont már évtizedek óta vannak érdemi gyakorlati eredmények. A cikkben bemutatott BME kutatási eredmények többek között a következő területeken kerültek alkalmazásra:

- a www.metnet.hu időjárásportál, illetve a Microsoft 2013-as fejlesztői versenyén nyertes Időjárás Mindenkinek Windows8 alkalmazás,
- számos MÁV-állomás hangos utastájékoztató rendszere,
- egy távközlési szolgáltató árlista bemondó szolgáltatása,
- egy távközlési szolgáltató automatizáltan kialakított interaktív hangválasz (IVR) rendszere,
- beszéd-dialógus mintarendszer intelligens lakás prototípusban a BelAmi projekt keretében,
- VoxAid2006 prototípus siketnéma emberek telefonálásának támogatására,
- VoxAid2012 prototípus beszédsérült emberek mindennapi kommunikációjának támogatására, logopédiai és afáziás betegek rehabilitációjának támogatására,
- prototípusrendszer mobil és autós információs szolgáltatásokhoz Android platformon,
- beszédvezérelt okostévé-készülék prototípus,
- beszélő mobil alkalmazások vak emberek számára Symbian, Windows Phone és Android platformon.

Köszönetnyilvánítás

Köszönöm elsősorban a BME TMIT Beszédkommunikáció és Intelligens Interakciók Laborcsoport csapatmunkáját, másrészt a BME TMIT munkatársainak, hallgatóimnak és kutatási partnereinknek az együttműködését. Az ábrák formázásában Németh Zsuzsanna volt segítségemre.

A cikkben áttekintett hazai kutatások eredménye többek között a BelAmi, 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 támogatásával valósultak meg).

Irodalom

Bánó Miklós. Tetszőleges szöveg reprodukálására alkalmas beszélőgép. Magyarország Szabadalom száma: 74361 . 1916. 11 30.

Baum, Leonard E – Ted Petrie 1966. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* 1554-1563.

Beutnagel, Mark – Conkie, Alistair – Jürgen, Schroeter – Stylianou, Yannis – Syrdal, Ann 1999. The AT&T next-gen TTS system. In Joint Meeting of ASA, EAA, and DAGA. 18–24.

Bőhm Tamás – Németh Géza 2006. Algoritmus formánsok követésére, módosítására és szintézisére. *Híradástechnika* LXI (8): 11–16.

Campbell, Murray – Hoane, A. Joseph Jr – Hsu, Feng-hsiung 2002. Deep Blue. *Artificial Intelligence* 57–83.

Cooper, Franklin S. 1961. Speech synthesizers. The Hague: Mouton & Co. Proceedings of the 4th International Congress of Phonetic Sciences, (Helsinki), 1961.

Cormen, Thomas H. – Leiserson, Charles E.– Rivest, Ronald L. 1990. Chapter 17 Greedy Algorithms. In *Introduction to Algorithms*, 768. Mcgraw-Hill.

Dahl, George E. – Dong, Yu – Deng, Li – Acero, Alex 2012. Context-Dependent Pre-Trained Deep Neural. *IEEE Transactions on Audio, Speech, and Language Processing* 20 (1): 30-42.

Dudley, Homer. – Riesz, R. R. – Watkins, S. A. 1939. A Synthetic Speaker. *J. Franklin Inst.* 227. 739–764. (Reprinted in Flanagan and Rabiner 1973).

Fék Márk – Pesti Péter – Németh Géza – Zainkó Csaba. Generációváltás a beszédszintézisben. *Híradástechnika* LXI (3): 21–30.

Flanagan, James – Rabiner, Lawrence (eds.) 1973. *Speech Synthesis.* Pennsylvania: Dowden, Hutchinson & Ross, Inc.

Furui, Sadaoki 2005. 50 Years of Progress in Speech and Speaker Recognition Research. *Ecti Transactions on Computer and Information Technology.* 64–74.

Gordos Géza – Takács György 1983. *Digitális beszédfeldolgozás.* Budapest: Műszaki Könyvkiadó.

Jelinek, Frederick 1976. Continuous speech recognition by statistical methods. *Proc. IEEE* 64: 532–536.

Kawai, Hisashi – Toda, Tomoki – Ni, Jinfu – Tsuzaki, Minoru – Keichi, Tokuda 2004. Ximera: a new TTS from ATR based on corpus-based technologies. *Proc. of the 5th ISCA Speech Synthesis Workshop.* Pittsburgh. 642–645.

Kempelen Farkas 1989. *Az emberi beszéd mechanizmusa, valamint a szerző beszélőgépének leírása.* Budapest: Szépirodalmi Könyvkiadó.

Kiss Gábor – Olaszy Gábor 1984. A Hungarovox magyar nyelvű, szótár nélküli, valós idejű párbeszédes beszédszintetizáló rendszer. *Információ Elektronika* 19 (2): 98–111.

Klatt, Dennis H – Klatt, Laura C. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America* 87 (2): 820–857.

Klatt, Dennis 1987. How Klattalk became DECtalk: An Academic's Experiences in the Business World. *Proc. of Speech Tech '87.* New York: Media Dimensions Inc. 293–294.

Lloyd, Richard J. 1890. *Some Researches into the Nature of the Vowel-Sound.* Liverpool: Turner and Dunnett.

Markov, Andrey A. 1913. 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*. 153–162.

McCulloch, Warren – Pitts, Walter H. 1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics* 5: 115–133.

Mermelstein, Paul 1973. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America* 53 (4): 1070–1082.

Mihajlik, Péter – Fegyó, Tibor – Tüske, Zoltán – Ircing, Pavel 2007. A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages – like Hungarian. *Proc. of Interspeech* 1497–1500.

Mori, Masahiro 970/2012. The uncanny valley. *(K. F. MacDorman & N. Kageki, Trans.). IEEE Robotics & Automation Magazine* 19(2), doi:10.1109/MRA.2012.2192811. 98–100.

Moulines, Eric – Charpentier, Francis 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communications* 9: 453–467.

Möbius, Bernd 2000. Corpus-based speech synthesis: methods and challenges. In Sendlmeier, Walter F. – Hess, Wolfgang (eds.): *Speech and Signals – Aspects of Speech Synthesis and Automatic Speech Recognition*. Frankfurt am Main. 79–96.

Nagy, Péter – Németh, Géza 2016. DNN-Based Duration Modeling for Synthesizing Short Sentences. *Proc. of Speech and Computer (SPECOM 2016).* Budapest: Springer International Publishing. 254–261.

Németh Géza – Olaszy Gábor 2010. *A magyar beszéd.* 1. Budapest: Akadémiai Kiadó.

Németh Géza – Olaszy Gábor – Fék Márk 2006. Új rendszerű, korpusz alapú gépi szövegfelolvasó fejlesztése és kísérleti eredményei. In *Beszédkutatás* 183–196.

Olaszy Gábor 1989. *Elektronikus beszédelőállítás. A magyar beszéd akusztikája és formánsszintézise.* Budapest: Műszaki Könyvkiadó.

Olaszy, Gábor 1989. MULTIVOX – A flexible text-to-speech system for Hungarian, Finnish, German, Esperanto, Italian and other languages for IBM-PC. *Proc. of Eurospeech '89.* Paris: European Speech Communication Association. 525–529.

Olaszy Gábor 1978. Szintetizált magyar magánhangzók formáns-intenzitás és formáns-sávszélesség értékei. *Magyar Fonetikai Füzetek* 68–77.

Olaszy, Gábor – Németh, Géza 1999. IVR for Banking and Residential Telephone Subscribers Using Stored Messages Combined with a New Number-to-Speech Synthesis Method. In Gardner-Bonneau, D. (ed.): *Human Factors and Voice Interactive Systems.* New York: Kluwer Academic Publishers. 237–256.

Olaszy, Gábor – Németh Géza, – Olaszi, Péter – Kiss, Géza – Zainkó, Csaba – Gordos, Géza 2000. Profivox– a Hungarian TTS System for Telecommunications Applications. *International Journal of Speech Technology* 3–4: 201–215.

Rosenberg, Aaron. G. – Schafer, Ronald. W. – Rabiner, Lawrence. R.1971. Effects of Smoothing and Quantizing the Parameters of Formant-Coded Voiced Speech. *J. Acoust. Soc. Am.* 50 (6B): 1532–1538.

Stevens, Kenneth N. – Kasowski, Stanley – Fant, C. Gunnar M.1953. An electrical analog of the vocal tract. *Journal of the Acoustical Society of America* 25 (4): 734–742. doi:10.1121/1.1907169.

Tokuda, Keichi – Yoshimura, Takayoshi – Masuko, Takashi – Kobayashi, Takao – Kitamura, Tadashi 2000. Speech parameter generation

algorithms for HMM-based speech synthesis. *Proc. of ICASSP.* Istanbul, Turkey: IEEE, 2000. 1315–1318.

Tóth, Bálint Pál – Németh, Géza 2008. Hidden Markov Model Based Speech Synthesis System in Hungarian. *Infocommunications Journal* LXIII (7): 30–34.

Tóth, Bálint Pál – Németh, Géza 2010. Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis. *Acta Cybernetica-Szeged* 19 (4): 715–731.

Van den Oord, Aaron – Kalchbrenner, N. – Kavukcuoglu, K. 2016. Pixel Recurrent Neural Networks. *arXiv preprint arXiv:1601.06759.*

Van den Oord, Aaron – Dieleman, Sander et al. 2016. A generative model for raw audio. *arXiv preprintarXiv:1609.03499.*

Waibel, Alex. – Hanazawa, Toshiyuki – Hinton, Geoffrey – Shikano, Kiyohiro – Lang, Kevin J.1989. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37 (3): 328–339.

Werbos, Paul 1974. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.* PhD thesis: Harvard University.

Zainkó Csaba –Tóth Bálint Pál – Németh Géza 2017. Magyar nyelvű WaveNet kísérletek. *Szegedi Egyetem.* Szeged: XIII. Magyar Számítógépes Nyelvészeti Konferencia. 205–216.

Zainkó, Csaba – Bartalis, Mátyás – Németh, Géza – Olaszy, Gábor 2015. A Polyglot Domain Optimised Text-To-Speech System for Railway Station Announcements. *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2015).* Dresden: International Speech Communication Association. 1236–1240.

Zen, Heiga – Senior, Andrew – Schuster, Mike 2013. Statistical Parametric Speech Synthesis Using Deep Neural Networks. *IEEE.* New York: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 7962–7966.

Zen, Heiga – Tokuda, Keichi – Black, Allan W. 2009. Statistical parametric speech synthesis. *Speech Communication* 51: 1039–1064.