MARKÓ ALEXANDRA – BEKE ANDRÁS

Beszél(get)ünk a számítógéppel? A beszéd mesterséges előállítása, számítógépes beszéd- és beszélőfelismerés

1. Bevezetés

A 21. század egyik új kutatási és alkalmazási területe a **beszédtechnológia.** A beszédtechnológia a mesterséges intelligencián belül a beszédalapú gyakorlati alkalmazások kifejlesztésével foglalkozik. A beszédtechnológia mindehhez a beszéddel kapcsolatos kutatások (például a fonetika), valamint az információtechnológia eredményeit használja fel (Németh–Olaszy szerk. 2010).

Jóllehet nem figyelünk rá, nem is vesszük észre, vagy legalábbis nem tudatosul bennünk, de mindannyian kommunikálunk számítógépekkel, és ez a kommunikáció sokszor beszéd segítségével zajlik. Gondoljunk arra, hogy amikor egy telefonközpont irányít minket a kívánt információhoz vagy a hívott félhez, az a telefonközpont egy „beszélő számítógép". Amikor a telefonszámlánk egyenlegét szeretnénk megtudni, azt akár egy számítógép bemondásában is meghallgathatjuk. Ha a tudakozót hívjuk fel, a kívánt telefonszámot ugyancsak egy számítógép közli velünk. De akkor is egy számítógép beszédére hagyatkozunk, amikor például az autóban navigációs rendszerre bízzuk a tájékozódást térképolvasás helyett. Arra is számos példát találhatunk, amikor a „gép leiratozza a beszédünket", mint amikor parancsszavakkal irányítunk valamilyen operációs rendszert, konzolt, vagy akár a név szerinti tárcsázást kívánjuk használni a telefonunkon. Vannak olyan diktáló rendszerek, ahol a gépnek felolvasott szöveg leiratozásra kerül (például orvosi leletezők). Ezeken kívül kereshetünk kulcsszavakat rövid hanganyagokban, híradós adásokban is. Az olvasók közül azonban a legtöbben valószínűleg a hangvezérelt okostelefonokkal, a videómegosztó webhely és a keresőfelület, illetve a fordító programok hangvezérelt keresésével találkozhattak.

Láthatjuk, hogy ahogyan az ember-ember közötti kommunikációt is beszédprodukcióra (a beszéd létrehozása) és beszédpercepcióra (a beszéd feldolgozása) osztja a tudomány, ugyanígy sorolhatók be az ember és gép közötti beszédkommunikáció alkalmazásai a számítógépes beszéd-előállítás, valamint a számítógépes beszédfelismerés nagy területeire. De természetesen – akárcsak az emberi kommunikáció, amelynek lényege a kölcsönösség – ezek a fejlesztések dialógusrendszerekben is működnek, vagyis vannak olyan alkalmazások, amelyekben az ember beszélve fordul a számítógéphez, amely ugyancsak beszéddel válaszol (pl. mesterséges intelligencia, robottechnológia, okosépületek).

DOI 10.18425/FONOLV.2016.17

A beszédtechnológia létrejötte és dinamikus fejlődése számos okra vezethető vissza. Ezek az alkalmazások részben a kényelmünket szolgálják, egyszerűsítik a munkát. Másrészt a számítógép sok műveletet gyorsabban végez el, mint az ember; nem téved; nem zavarja a monotónia (vagyis ugyanazt a feladatot akár több tízezerszer is el tudja végezni egymás után ugyanúgy, anélkül, hogy belefáradna, és nem hibázik); ötvözni tudja többféle alkalmazás előnyeit (például a navigációs rendszer a helymeghatározást és a térbeli tájékozódást az információközléssel). A beszédtechnológiai alkalmazások fontos előnye, hogy a valamilyen sajátos igénnyel élő emberek számára nagy segítséget nyújtanak. Például meghangosítják a számítógép képernyőjét a vakok és gyengénlátók számára; súlyos beszédzavar esetén közvetítik a beszélő szándékát; a siketek és nagyothallók számára a beszédjelet a száj mozgását ábrázoló vizuális jellé alakítják.

A következőkben néhány olyan alkalmazást és ezek működési módját mutatjuk be röviden, amelyeknél az ember és a számítógép közötti kommunikáció beszéd segítségével zajlik.

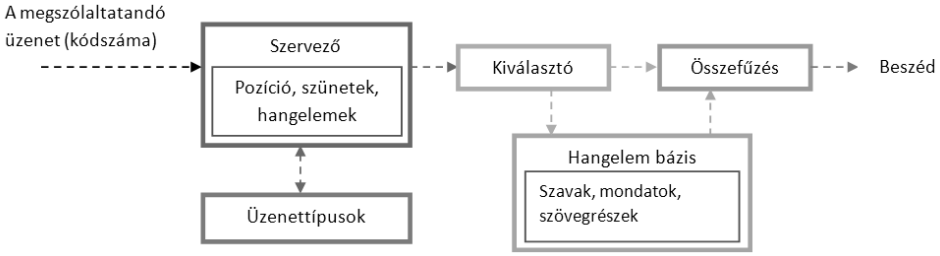
2. Számítógépes beszéd-előállítás

A számítógépes beszéd-előállításon – más szóval: beszédszintézisen – belül az alkalmazások két nagy csoportját különíthetjük el: az ún. kötött szótáras rendszereket és a szövegfelolvasókat. A kötött szótáras rendszerek zárt (és általában nem túl nagy) szókészlettel működnek, néhány üzenettípust tudnak meghangosítani, amelyeknek a szerkezete is kötött. A szövegfelolvasók ezzel szemben (elvben) bármilyen üzenet meghangosítására képes, rugalmas rendszerek.

2.1. Kötött szótáras rendszerek

A kötött szótáras rendszerek olyan alkalmazások, amelyekkel nap mint nap találkozunk, ha tömegközlekedési eszközökön utazunk, GPS-es útvonaltervezést használunk, hivatalokkal vagy vállalatokkal kell telefonhívást lebonyolítanunk, vagy olyan kényelmi szolgáltatásokat veszünk igénybe, mint a telefonébresztő vagy a telebank. Ezek a rendszerek emberi beszédből előre felvett rövid szövegrészleteket (egyedi beszédelemeket, vö. Olaszy 1999) fűznek össze egy meghatározott logika szerint. Ezekben a felhasznált szövegrészletek száma nem túlságosan nagy (néhány tíz), és ugyancsak nem túl nagyszámú kombinációkban hangoznak el. Gondoljunk például egy számítógép vezérelte telefonközpontra, vagy akár a járműveken szóló utastájékoztatókra! Ezeknek a létrehozásához először megtervezik azt a nyelvi anyagot, amelyet rögzíteni érdemes. Az adott vállalat kiválaszt egy beszélőt, akinek a bemondásában felveszik ezeket a részleteket. Az így összeállított hangelemtárból fogja az összefűző algoritmus kiválasztani a szükséges elemet. Az ismétlődő része-

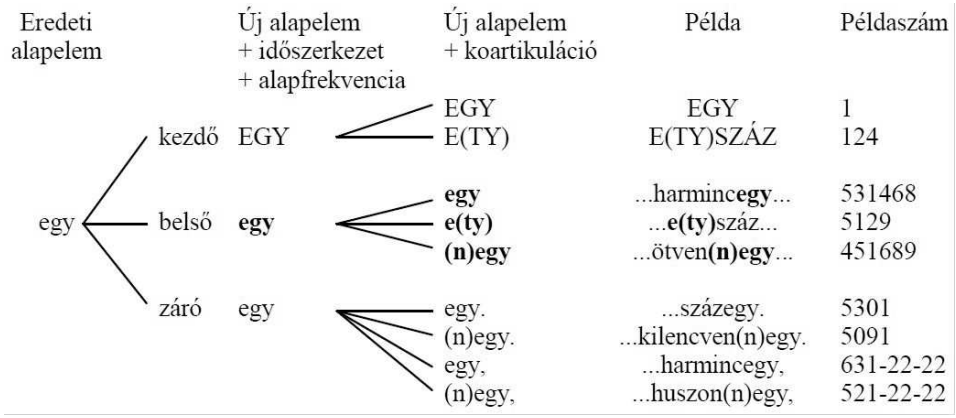
ket csak egyszer veszik fel, de az algoritmus több különböző körben többször ki tudja ezeket választani. Az 1. ábra mutatja be a kötött szótáras szintézis folyamatát és szükséges moduljait. A hangelembázis az üzenetek, üzenetrészek hullámformáit tárolja. A kiválasztó algoritmus meghatározza, hogy az elemtárból mely összefűzendő hangelemeket (üzenetrészeket) kell kiválogatni a bejövő parancs alapján, majd az összefűző modul egymáshoz kapcsolja ezeket.



1. ábra. Az általános kötött szótáras beszédszintetizátor blokksémája (Forrás: Pandur 2011: 10)

Ha a kötött szótáras rendszer csak változatlan mondatokat tartalmaz (pl. *A Kálvin-tér következik. – Kálvin tér. – Az Astoria következik. – Astoria.*), akkor a hangzása jó minőségű lesz, hiszen csak az ember által bemondott szerkezeteket szólaltatja meg változtatás nélkül.

Ha az alkalmazás változó elemeket is tartalmaz, a hangzás minősége nagyban függ attól, hogy a tervezéskor mennyire jártak el körültekintően. Például egy egyenlegközlő vagy az üzenetrögzítőnk lehallgatásakor értelemzavaró lehet, ha a számokat nem olyan bontásban hangosítja meg a rendszer, amelyhez szokva vagyunk; vagy ha a számkapcsolatok hangzása nehezíti a megértést. Képzeljük el, hogy a bankszámlánk egyenlegét szeretnénk lekérdezni, amely 25 974 forint, és a gép a következő információt közli, a | helyén rendre szünetet tartva: *Az ön egyenlege* | *huszon* | *öt* | *ezer* | *kilenc* | *száz* | *hetven* | *négy* | *forint.* A bemondás még nehezebben értelmezhető, ha a gép nemcsak szünetet tart az elemek között, de mindegyiket külön is „hangsúlyozza". Ez könnyen megtörténhet, ha a számelemek felvételekor nem figyeltek az összetartozó számelemek hangzására, csak felvették a magyar nyelvű számalakok főbb elemeit külön-külön. Ahhoz, hogy jól hangzó számfelolvasót állítsunk elő, többek között olyan fonetikai ismeretekkel kell rendelkeznünk, mint hogy milyen a különféle magyar szerkezetek hangsúlyozása, hanglejtése; a beszédhangok hogyan hatnak egymásra; hol tarthatunk szünetet, és hol lenne értelemzavaró a szünettartás, stb. A 2. ábra arra mutat példát, hogy mindezek az ismeretek hogyan hasznosulnak egy számfelolvasó kötött szótáras rendszer tervezésekor.



2. ábra. Az *egy* számelem optimális változatai egy fonetikailag jól megtervezett számfelolvasó esetén (Forrás: Olaszy 2010a: 288)

Az ábrán az *egy* számelem különféle szükséges bemondásait látjuk. A tervezéskor figyelembe vették, hogy az *egy* elem kezdő (pl. ***egy****ezer*...), belső (pl. *kétezer-****egy****százhat*) vagy záró helyzetű (pl. *négyszázkilencven****egy***) is lehet. A kezdő helyzetben az *egy* főhangsúlyosan kell, hogy elhangozzon, magasról induló dallammal; míg a záró helyzetben hangsúlytalanul, alacsony hangmagasságon. Az *egy* esetében a mássalhangzó hosszúsága is függ a számelem helyzetétől, illetve az őt követő számelem kezdő hangzójának minőségétől. Ha az *egy* szerkezet végén áll, a *gy*-t hosszan ejtjük, akárcsak két magánhangzó között, mint például az *egyezer* esetében. Ha azonban kezdő vagy belső helyzetű, és mássalhangzó követi (pl. *egymillió*), a *gy* ejtése rövid lesz. Mint a példából látható, a jó minőségű hangzás szempontjából az sem mindegy, hogy figyelembe vesszük-e, milyen számelem követi az épp aktuálisat. Az *egy* olyan zöngés mássalhangzóra végződik, amely részt vesz a zöngésségi hasonulás fonológiai folyamatában. Vagyis ha ezt a számelemet zöngétlen mássalhangzó követi, az *egy gy*-je *ty*-vé zöngétlenül. Ez történik az *egyszáz* → *etyszáz* kiejtésekor. A modell azt is figyelembe veszi, ha a belső vagy záró helyzetben álló *egy* előtt nazális mássalhangzó van, hiszen ebben az esetben az *egy* magánhangzója nazalizálódhat. Mindezek alapján a jól megtervezett számfelolvasóban az *egy*-nek 9 változatát kell rögzíteni és eltárolni.

2.2. Automatikus szövegfelolvasás

A szövegfelolvasók (angol rövidítésük alapján TTS-nek is nevezik, mint text-to-speech, szöveg-beszéd átalakító) a kötött szótárasoknál sokkal bonyolultabb rendszerek. Hiszen míg a kötött szótáras alkalmazások előre megadott szövegelemeket használnak, addig a szövegfelolvasóknak az a céljuk, hogy (elvben) bármilyen témájú, műfajú szöveget képesek legyenek az emberi beszédhez hasonló hangzással

meghangosítani. Ahhoz, hogy ezt a célt el tudják érni, nyelvi és prozódiai modelleket és elemzőket használnak, amelyek a megszólaltatáshoz szükséges információkat kinyerik az írott szövegből (Olaszy 2010b). Például megadják, hogy a gépi felolvasó hol tartson szünetet, hová tegyen hangsúlyt stb. Egyes szintézismódszerekben ezeknek nagyobb a jelentőségük, másokban kisebb.

Ahhoz, hogy a számítógép meg tudja hangosítani az írott szöveget, mindenekelőtt olyan formára kell hozni az írást, hogy azt a gép fel tudja dolgozni, és abból hangzó anyagot tudjon előállítani. Ehhez olyan módosításokra van szükség, mint például a számoknak kiejthető betűsorrá való átalakítása (pl. *8.25 → nyolc óra huszonöt perc, 1945. 05. 09. → ezerkilencszáznegyvenöt május kilenc*). A betűszavakat, rövidítéseket, mértékegységeket fel kell oldani (pl. *stb. → satöbbi, kg → kilógram, SMS → esemes*), az idegen vagy hagyományos írásmódú szavakat ugyancsak a számítógép által feldolgozható, egységes írásmódra kell alakítani (pl. *e-mail → ímél, New York → nyújork, Batthyány → battyányi*). SMS-ek, e-mailek estében az is gyakori, hogy a karakterek számának csökkentése céljából a szöveg írója betűszám kombinációkat vagy nem szokványos rövidítéseket alkalmaz, ezeket is normál formára kell hozni (pl. *5let → ötlet, Lmegy → elmegy, vok → vagyok*). Az emotikonok kezelése is az előfeldolgozás része.

Ha a szöveg már a kívánt betűkarakterek sorozataként áll rendelkezésre, ezt követően lehet létrehozni a kiejtési modellt a magyar nyelv fonológiai-fonetikai szabályai alapján. Vagyis a betűsort beszédhangokká alakítják. Itt azokat az ismereteket kell alkalmazni, hogy például mi történik két mássalhangzó vagy két magánhangzó találkozásakor a folyamatos beszédben (pl. az előbbi esetében hasonulások, rövidülések, kiesések, az utóbbi estében hiátustöltés jelentkezhetnek). Ezeket általában szabályok vezérlik, hiszen zöngés és zöngétlen mássalhangzók találkozásakor mindig történik zöngésségi hasonulás. Ugyanakkor szükség van egy ún. kivételszótárra is, amely azokat az eseteket tartalmazza, amikor valamilyen elvárt alkalmazkodási folyamat mégsem történik meg. Például az *átjön* szót nem ejtjük *tty*-vel, pedig a *tj* kapcsolatokat általában igen *(botja, látja, mutatja)*.

A nyelvi elemzés és átalakítás után kerül sor az előállított beszéd meghangosítására, vagyis az akusztikai szerkezet megvalósítására. A fejlesztők erre többféle módszert dolgoztak és dolgoznak ki, és ezek összekapcsolása (ún. hibrid módszerek fejlesztése) is gyakori megoldás. A létező módszertanok közül most csak a formánsszintézist, a diád-triád alapú szintézist és az ún. korpuszos (változó elemhosszúságú) szintézist mutatjuk be. Ezek a beszéd akusztikai szerkezetéből indulnak ki, de vannak olyan módszerek is, amelyek az emberi beszédkeltés artikulációs sajátosságait modellezik és másolják le. (A további tájékozódáshoz ajánljuk a Németh és Olaszy által szerkesztett kötetet (2010) és a hozzá tartozó honlapot: http://magyarbeszed.tmit.bme.hu/index.php?p=home.)

A **formánsszintézis** a számítógépes beszéd-előállításra elsőként kidolgozott módszer. Az USA és Svédország jártak legelöl a módszer alkalmazásában, és Magyarországon volt a harmadik olyan kutatóhely, ahol létrehoztak ilyen rendszert.

A HUNGAROVOX-ot 1982-ben mutatták be. A formánsszintézis módszer lényege, hogy a számítógéppel leutánozzák az emberi beszédet, oly módon, hogy az emberi beszédből elemzett és kivont sok-sok paramétert betáplálják a számítógépbe egy beszédmodell segítségével. A beszédet előállító szabályok összeállítása hosszadalmas, és mély fonetikai szaktudást igényel. A legismertebb és legszélesebb körben alkalmazott magyar formánsszintetizátor a MULTIVOX (1990 óta), amely szabadon hozzáférhető és használható (http://magyarbeszed.tmit.bme.hu/index.php?p=multivox_letoltes). A honlapon megszólaltatható hangminták alapján jól hallható, hogy ennek a hangzása robotos, gépies, fémes, tehát nem közelíti meg az emberi hangszínezetet, ugyanakkor nagyon jól érthető. A formánsszintézisnek nagy előnye, hogy kis tárkapacitást igényel, ezért bármilyen (akár régi) számítástechnikai eszközön, telefonon futtatható. Mivel (hozzáértéssel) nagyon könnyen lehet módosítani a hangot, jól alkalmazható olyan fonetikai kísérletekben, amelyeknél egy-egy paraméter szisztematikus módosítása a cél (amire egy humán beszélő nem volna képes).

A számítógépes beszéd-előállításban ma már sokkal jellemzőbb, hogy emberi beszédből hoznak létre adatbázisokat, és a szövegfelolvasók ezekből állítják össze a hangzó beszédet. Hogyan történik mindez?

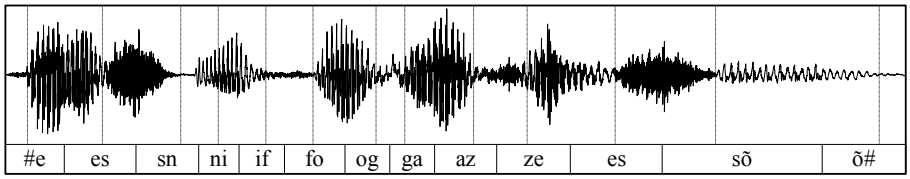
Egy fonetikában járatlan személy előállhatna egy olyan ötlettel, hogy állítsunk elő magyar beszédet a számítógép segítségével úgy, hogy felvesszük az összes (a magyar esetében kb. 40 db) beszédhangot egy valaki bemondásában, és létrehozunk egy algoritmust, amelynek nem volna más dolga, mint hogy ezeket az elemeket az írott forma alapján egymás után, szünet nélkül lejátssza. Milyen lenne ez a beszéd? Természetellenes lenne a ritmusa, a hangsúlyozása, a dallama, a tagolása. Valószínűleg annyira eltérne az emberi beszédtől, hogy első hallásra meg sem értenénk, de biztosan nagyon fárasztó lenne hallgatni.

Egy fonetikai ismeretekkel rendelkező szakembernek azonban vannak ismeretei arról, hogy a beszédhangokat folyamatosan formáljuk, és a közöttük lévő hangátmenetek igen fontosak mind az artikuláció, mind pedig a beszédfeldolgozás szempontjából (lásd Gósy Mária (2016) írását a jelen kötetben). Valamint azt is tudja, hogy milyen nagy a beszédben a szupraszegmentumok jelentősége. Ezért ő feltehetőleg módosítaná az ötletet. Két javaslata lenne: 1. ne a beszédhangokat vegyük fel önmagukban, hanem rögzítsük a hangátmeneteket; 2. hozzunk létre egy prozódiai modellt, amely gondoskodik arról, hogy a beszéd szupraszegmentális szerkezete megfelelő legyen!

Ezzel a logikával hozták létre az ún. **diád**os szintézist. A diádok két fél beszédhangnyi hullámformarészletet tartalmaznak (lásd 3. ábra). Ha azt szeretnénk, hogy a számítógép a *Jó napot!* hangsort mondja ki, ehhez a következő diádokra lesz szükségünk:

1. diád: [szünet] + a *j* beszédhang első fele,
2. diád: a *j* beszédhang második fele + az *ó* beszédhang első fele,
3. diád: az *ó* beszédhang második fele + a *n* beszédhang első fele,
4. diád: a *n* beszédhang második fele + az *a* beszédhang első fele,

5. diád: az *a* beszédhang második fele + a *p* beszédhang első fele
és így tovább.

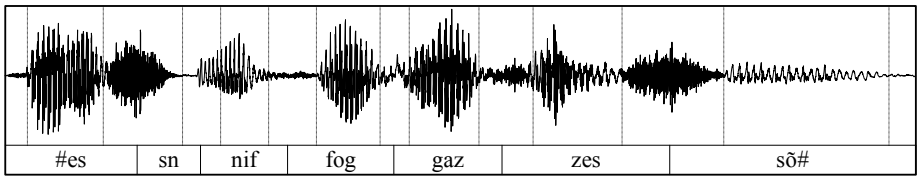


3. ábra. Az *Esni fog az eső mondat* szintetizált hullámformája 13 diádból összefűzve (Forrás: Olaszy 2010: 293)

40 beszédhanggal számolva 1600 diád nagyjából lefedi az egy nyelv hangzásában megjelenhető beszédhangkapcsolatokat. Ez azért előnyös, mert viszonylag kisméretű adatbázist kell kezelni. Ugyanakkor nem mindegy, hogy ezeket a diádokat hogyan vesszük fel, ez gondos tervezést igényel. Például a diádokat értelmetlen hordozó hangsorokban (álszavakban, logatomokban), monoton kiejtéssel kell felvenni, állandó tempóval, mert csak így lehet biztosítani, hogy a hangzásuk igen hasonló legyen. A diádos adatbázisból behívott, majd egymás után fűzött elemek azonban önmagukban még nem adnak jól hangzó beszédet. Ehhez szükséges a megfelelő prozódia ráültetése, az akusztikai paraméterek módosításával. Ilyen például az, hogy eldöntendő kérdő mondat esetében az utolsó előtti szótagon frekvenciacsúcsot kell elhelyezni (ennek akusztikai paraméterei természetesen aprólékos tervezést kívánnak meg), vagy az, hogy vesszőnél szünetet tart, és új dallamívet indít a szövegfelolvasó.

Ilyen diádos módszerrel hozták létre a magyar ProfiVox (1995 óta) beszédszintetizátor egyik első változatát (http://magyarbeszed.tmit.bme.hu/index.php?p=profivox_letoltes).

A diádos szintézisnek sok előnye (az adatbázis kis tárigénye, egyszerű kezelhetősége, kis hibaaránya, könnyű javíthatósága) mellett hallható hátránya volt, hogy az elemek összefűzési helyein az illesztés gyakran hallatszott, és ez torz, gépies hangzást idézett elő, még annak ellenére is, hogy az adatbázist emberi beszédből hozták létre. Ennek kiküszöbölésére merült fel az az ötlet, hogy ne legyen illesztés nagy energiájú helyeken (ahol ez a legjobban hallatszik), azaz a magánhangzókat ne két elemből illesszék össze. Ennek érdekében a ProfiVox adatbázisát ún. **triád**okkal egészítették ki (vö. 4. ábra), ezeknek a szerkezete „mássalhangzó második fele + magánhangzó + mássalhangzó első fele" volt. A magánhangzó-magánhangzó és a mássalhangzó-mássalhangzó kapcsolatokra továbbra is diád elemeket használtak.



4. ábra. Az *Esni fog az eső mondat* szintetizált hullámformája 6 triádból és 1 diádból összefűzve (Forrás: Olaszy 2010a: 299)

A számítógépes technológia fejlődésével egyre kevésbé volt annak jelentősége, hogy a beszédszintézishez használt adatbázisok kis tárhelyet foglaljanak. Ennek megfelelően a magyar fejlesztések is a nagy beszédadatbázisok irányában indultak el, elsődlegesen a változó elemhosszúságú egységek kiválasztásának módszertana, később pedig a gépitanulás-alapú módszerek felé. Összefoglalónkban az előbbit tárgyaljuk részletesen.

Az **elemkiválasztásos technológia** a beszédadatbázisban (korpuszban – ezért gyakran korpuszos szintézisnek is nevezik) való közvetlen keresésen és a talált hullámformarészek közvetlen összefűzésén alapul (Olaszy 2010a: 300). Az adatbázis ez esetben többórányi, célzottan felcímkézett beszédet tartalmaz (lásd Varjasi Gergely (2016) írását a jelen kötetben). Ellentétben a diád-triád alapú adatbázisokkal, ez bővíthető, és nem kötött elemeket, hanem mondatokat tartalmaz. A bemondó szempontjából ez azt jelenti, hogy míg a diádok és triádok felvétele néhány órás igénybevétel, a korpuszos szintézishez több ülésben, alkalmanként több órán keresztül kell felvenni a hanganyagot, ráadásul úgy, hogy a beszélő hangszínezete eközben, illetve alkalomról alkalomra nem változhat jelentősen. Hiszen ha így lenne, az a szintézisben komoly minőségi romlást okozna (akár szavanként, szintagmánként eltérő hangzást). Ez egyrészt azt jelenti, hogy a bemondónak professzionális beszélőnek kell lennie, másrészt azt is, hogy egy adatbázishoz csak egy személy hangja használható fel.

Míg a diádos-triádos vagy a formánsszintézissel bármilyen szöveg meghangosítható, a korpuszos szintézis témaspecifikus. Ahhoz ugyanis, hogy jól hangzó beszédet valósíthassunk meg vele, valamilyen témára kell korlátoznunk a felolvasható szöveget. Mindenki számára elérhető a metnet.hu oldalon működő elemkiválasztásos szintézis, amely a napi időjárás-előrejelzést hangosítja meg, de ezzel a módszerrel működik a Keleti pályaudvar új utastájékoztató rendszere is (mindkettő a BME TMIT fejlesztése).

Az elemkiválasztásos szintézis előállítása röviden: felmerül az igény valamilyen témában arra, hogy a számítógép hangosítson meg szövegeket. A fejlesztők az adott témában gyakran elhangzó közléseket összegyűjtik (pl. kivonatolnak időjárásjelentéseket), és különféle szempontok alapján összeállítanak 5-10 ezer mondatot. A szempontok között szerepel például, hogy a gyakori kifejezések a mondatban többféle helyzetben jelenjenek meg (a mondat elején, közepén, végén, tagmondat-

határon stb.). Erre azért van szükség, mert az elemkiválasztásos módszerben nincs külön prozódiaráültetés (vagyis nem utólag kapja meg a szöveg a hangsúly-, hanglejtés- stb. mintázatokat), hanem már az elem kiválasztásakor igyekszik a válogató algoritmus olyan részletet találni, amely a mondatbeli helyzetét tekintve is hasonlít a felolvasandó szerkezet mondatbeli helyzetéhez.

Nézzünk erre egy példát! A szintetizátornak a következő mondatot kell meghangosítania: *A szélcsendes délnyugati völgyekben néhol ködfoltok képződhetnek.* A válogató algoritmus a lehető legnagyobb mértékben egyező részt keresi az adatbázisban. Ha talál olyan mondatot, amely teljes egészében megfelel a felolvasandónak, akkor azt választja be. Ha ilyet nem talál, akkor próbál minél nagyobb egyezést találni. Feltételezhetjük, hogy megtalálja a *néhol ködfoltok képződhetnek* szerkezetet. Önmagában a szöveg szerinti egyezés azonban nem elegendő, mert ha ezt az egybefüggő részt egy mondat elején találja, a *néhol* sokkal magasabb dallammal indulna ahhoz képest, mint amilyet a fenti, előállítandó mondatban elvárnánk, ahol a mondat végén szerepel ez a szerkezet. Így nem választja ki az algoritmus ezt a szerkezetet, hanem tovább keres, míg prozódiailag megfelelőt nem talál. Ha nem talál megfelelőt, akkor a *néhol,* a *ködfoltok* és a *képződhetnek* szavakat esetleg külön-külön fogja kiválasztani az adatbázisból, olyan helyekről, ahol ezek mondat belsejében vagy végén szerepelnek.

Természetesen így is előfordulhat, hogy az algoritmus csak olyan elemeket talál az adatbázisban, amelyek nem felelnek meg prozódiai szempontból. Ekkor egy ilyet fog kiválasztani, a fejlesztők pedig a minőség-ellenőrzés során – hallva az eltérést – korrigálják ezt a hibát oly módon, hogy az adatbázist bővítik az adott szerkezetet megfelelő prozódiával tartalmazó mondattal. (Az ilyen hibákat folyamatosan gyűjtik, és időről időre újabb megtervezett mondatcsoportokat olvastatnak fel a bemondóval, akinek ezen alkalmakkor újra rá kell hangolódnia a korábbi beszédminták beszédsajátosságaira.)

Mindemellett az is lehetséges, hogy egy-egy kisebb-nagyobb nyelvi egység nem áll rendelkezésre az adatbázisban. Ekkor a ProfiVox a diád-triádos elemtárból pótolja ki a hiányt.

A gépi tanulásos módszerek (ma Magyarországon ez a rejtett Markov-modell alapú, HMM-szintézist jelenti) ugyanilyen aprólékosan címkézett beszédadatbázisokat használnak. A gépi tanulás előnye egyebek mellett az, hogy kisebb induló adatbázis is elegendő, mivel a gépi tanuló algoritmusok fel tudják venni a bemondó beszédsajátosságait, így a bővítéshez sem szükséges a bemondó jelenléte. Ebből az is következik, hogy sem a téma, sem a bemondás hangzása nem korlátozott, bármilyen beszélő hangjára lehet adaptálni ezeket a rendszereket. Ezeknek a fejlesztése előtt azonban még hosszabb út áll.

3. Számítógépes beszédfelismerés

Az automatikus beszédfelismerés szintén egy nagyobb csoportot alkot, hiszen számos részterületet foglal magában. A beszédfelismerés legismertebb célja, hogy az ember által kimondott szavakat, szövegeket automatikusan leiratozza, vagyis az ember által gerjesztett hullámformát szöveges karakterekké alakítsa. Emellett azonban számos más részterület is ide tartozik, hiszen a tartalmi leiratozáson kívül az is fontos lehet, hogy kitől származik az elhangzott beszéd, milyen érzelmi, egészségügyi állapotban van az illető. A „Ki beszél?" kérdéssel az automatikus beszélőfelismerés foglalkozik. Az egyén hangulatának automatikus felismerésével az érzelemfelismerés, az egészségügyi állapot meghatározásával pedig a klinikai beszédfelismerés.

3.1. Beszédfelismerés

A beszéd komplex folyamat, ahol az információ akusztikai formában közvetítődik, azonban nyelvi tartalmat hordoz. Ezért a beszédfelismerésben pusztán csak az akusztikai jel felől közelíteni nem elégséges, hanem valamiféle nyelvtant is létre kell hoznunk, amely azt adja meg, hogy az egyes modellezett beszédegységek (lehetnek ezek beszédhangok, szótagok, szavak stb.) milyen valószerűséggel követik egymást. A beszédfelismerő tehát két nagyobb részből tevődik össze, egy akusztikai modellezésből és egy nyelvtanból. A mai beszédfelismerők szinte kizárólag rejtett Markov-modellt használnak.

A gépi beszédfelismerésnek többféle változata van az artikuláció, a beszélő, az akusztikai környezet és a szótárméret függvényében. Az artikuláció szerint lehet izolált szavas beszédfelismerő, amely szavak felismerésére alkalmas, illetve folyamatos, amely képes folyamatos beszéd felismerésére, így ez áll a legközelebb az emberi beszédfelismeréshez. A beszédfelismerő lehet beszélőfüggő, illetve beszélőfüggetlen. A beszélőfüggetlen felismerő alapvető célja, hogy olyan modelleket alkosson, hogy a beszédfelismerőt bármely felhasználó használhassa. Az akusztikai környezet szintén jelentősen befolyásolja a beszédfelismerő működését. Csendes körülmények között elhangzott beszéden a beszédfelismerő pontos eredményt tud adni, míg zajban az eredmények jelentősen romlanak. Kiemelt feladat a telefonos beszélgetések automatikus leiratozása. Ez a feladat abban tér el a fentiektől, hogy a telefon (telefontípustól függően) más-más frekvenciasávot ereszt át (jellegzetesen 300–3800 Hz közötti frekvenciatartományt). A beszédfelismerésben használt szótár mérete is fontos szempont. Léteznek kicsi (< 100 szó), közepes (100–1000 szó), nagy (> 10 000 szó) és kötetlen szótáras beszédfelismerők. Ez azt adja meg, hogy hány szó felismerésére képes a rendszer.

A beszédfelismerő rendszer eredményességét számos körülmény nehezítheti. Az egyik ilyen a beszédstílus, hiszen a felolvasott beszédet a rendszer közel 80-90%-os pontossággal ismeri fel, a spontán beszédet csupán 50-60%-os pontossággal, hiszen

ennek akusztikai változatossága nagyobb, a beszédhangok kiejtése pontatlanabb, illetve a nyelvtana is kevésbé feltérképezett. A másik nehézség a nyelvi adaptálhatóság. A morfológiailag gazdag nyelvekre, mint amilyen a magyar is, a felismerés eredménye rosszabb, mivel igen nagy a ritka szavak száma (itt a *kutya, kutyának, kutyái* stb. mind külön szóként reprezentálódik a szótárban), ezért nagyon nagy szótárra lenne szükség, illetve igen nagy a szótáron kívüli elemek száma is.

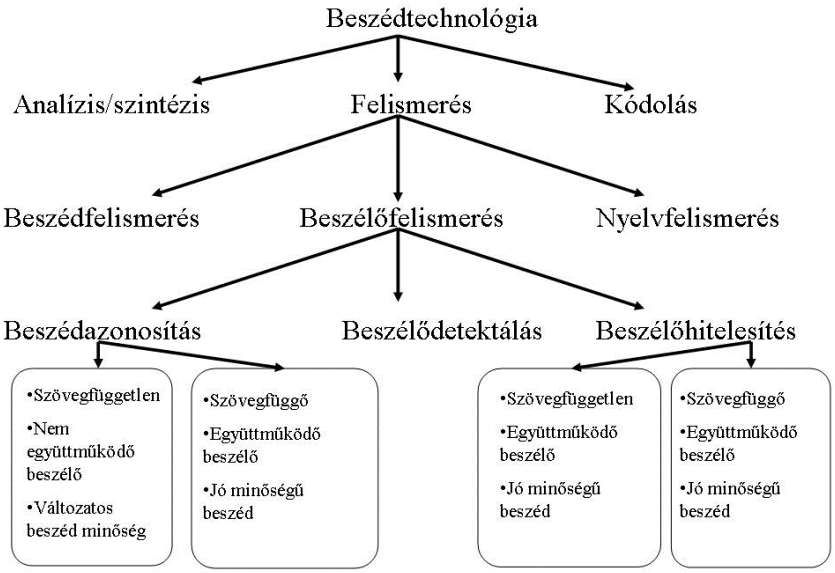
A beszédfelismerő rendszert számos helyen alkalmazzák a már eddig említetteken kívül, mint az orvosi lelet automatikus leiratozása vagy a beszédterápia, az olvasásfejlesztés (pl. Beszédmester). További alkalmazás az audiovizuális számítógépes beszédfejlesztő program beszédhibás gyermekek részére (Varázsdoboz).

A beszédfelismerésről részletesebb leírás olvasható (Németh–Olaszy szerk. 2010; Mihajlik 2013).

3.2. Beszélőfelismerés

A mindennapi életben képesek vagyunk akár néhány másodperces hangmintából azonosítani az általunk ismert személyeket. Ez azért lehetséges, mert a beszédhang olyan akusztikai jegyeket tartalmaz, amelyek jól reprezentálják az adott egyént (Bőhm 2007). Kutatások kimutatták, hogy a hangfelismerésért, akárcsak az arcfelismerésért, egy külön agyi terület felelős. Képalkotó eljárások ugyanis bizonyították, hogy más-más agyterület aktiválódott az ismert és nem ismert személy beszédének hallgatása során (Belin et al. 2004, idézi Bőhm 2007). Az ismert személyek felismerése mellett képesek vagyunk a nem ismert személyekről is profilt készíteni, vagyis általános információkat adni például a nemre (Lass et al. 1976), az életkorra (Ptacek–Sander 1966; Gocsál 1998), testalkatra (Dommelen–Moxness 1995; Gósy 2001) vagy hangulatra (Scherer–Banse–Wallnott 2001) stb. vonatkozóan.

A gépi beszélőfelismerés alapvetően három területre osztható (vö. 5. ábra). Megkülönböztetünk beszélőazonosítást (speaker identification), beszélőhitelesítést (speaker verification) és beszélődetektálást (speaker diarization) (Bimbot et al. 2004). A beszélőhitelesítés célja, hogy a rendszer egy személyről eldöntse, hogy ő az, akinek állítja magát. Ez a cél megegyezik a többi biometrikus személyazonosítás (pl. ujjlenyomat, íriszvizsgálat) céljával. Ebben a feladatban bináris döntést kell hoznia a gépnek: elfogadás/elutasítás. Ekkor a beszélőnek érdeke, hogy a gép felismerje a hangját, ezért a beszédminőség igen jó. Ezzel ellentétben a beszélőazonosítás célja, hogy a beszélők egy lehetséges köréből kiválasszuk az aktuálisan beszélőt. Lehetséges azonban az is, hogy a lehetséges beszélők halmaza nyílt, vagyis a beszélő nincs benne a halmazban, ekkor a rendszer ismeretlen személyként kell, hogy azonosítsa a beszélőt. A beszélődetektáláskor két- vagy több-beszélős társalgásokban kell azonosítani azt, hogy ki mikor beszél. A beszélőazonosítás és a beszélőhitelesítés lehet szövegfüggő vagy szövegfüggetlen. A kutatók általában a szövegfüggetlen osztályozásra törekszenek, mivel ekkor tetszőleges tartalmú beszédminta alapján történhet a beszélő azonosítása vagy hitelesítése.



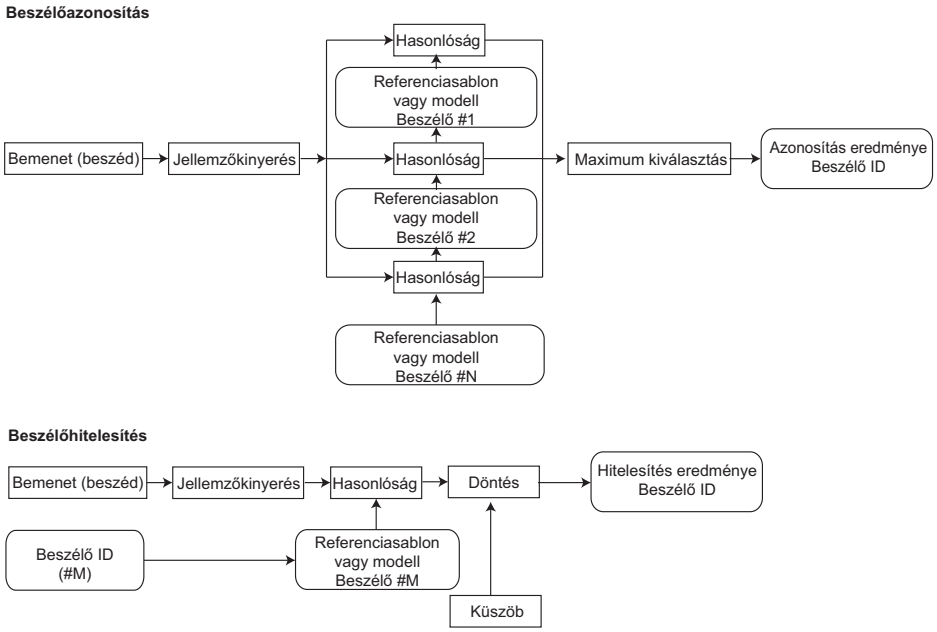
5. ábra. A gépi beszédfeldolgozás területei

A beszélőhitelesítés napjainkban egyre inkább megoldottnak tűnik, mivel közel 98-99%-os eredménnyel működik. A beszélőazonosítás eredményei ehhez képest jóval változatosabbak. Az eredmények nagyban függnek a felvétel minőségétől, azaz hogy milyen zajos a felvétel, és a beszédminta hosszától stb. A gyakorlatban legtöbbször igen rövid akusztikailag feldolgozható minta áll rendelkezésre az azonosításhoz (Nikléczy 2001, Nikléczy–Gósy 2008). Kutatások szerint a legrövidebb beszédminta hossza, ami még alkalmas az azonosításra, 16 másodperc (Nikléczy–Gósy 2008).

A beszélőfelismerés öt lépésből áll: a beszédjel tisztítása, jellemzőkinyerés, beszélőmodellek létrehozása, mintaillesztés, döntés (6. ábra).

A bemeneti beszédjelből eltávolítjuk azokat a részeket, amelyek nem járulnak hozzá a beszélő személy felismeréséhez vagy nehezítik azt. Ilyen tipikus eljárás a zajszűrés, beszédjeltisztítás, amely során a beszédből eltávolítjuk a zaj minél nagyobb részét, javítva ezzel a jel/zaj viszonyt. A másik eljárás a beszéddetektálás, amely során csak azokat a részeket tároljuk el, ahol a beszélő valóban beszél, kiszűrve ezzel a szüneteket, hosszabb légvételeket, zajos részeket. A beszédjel megtisztítása után számítjuk ki az akusztikai jellemzőket. Az akusztikai jellemzők igen sokfélék lehetnek. A jellemzőkinyerés célja az, hogy megtaláljuk azon akusztikai tulajdonságokat, amelyek mentén az egyes beszélők elkülöníthetők, azaz amelyek beszélőszemély-specifikusak. Az akusztikai jellemzőknek ugyanakkor egyszerűen mérhetőnek, minden beszélőnél jól mérhetőnek, érzelmi állapottól függetlennek kell

lenniük. A feladatra használt akusztikai jellemzők száma igen nagy, azonban továbbra is kérdés marad, hogy létezik-e, és ha igen, akkor mely akusztikai paraméterben mutatható ki az egyéni hangszínezet.



6. ábra. A beszélőazonosítás (fent) és a beszélőhitelesítés (lent) folyamatábrája

A jellemzőkinyerés után előállnak az úgynevezett jellemzővektorok, amelyek alapján elvégezhető az osztályozás. Az osztályozáshoz a beszédfelismerésben is használt algoritmusokat szokás alkalmazni (pl. kevert Gauss-modell, rejtett Markov-modell, neurális hálózatok, szupport vektor gépek és ezek kombinációi). A beszédfelismeréshez képest azonban a beszélőhitelesítéskor (6. ábra lent) a modellek közötti hasonlóság mérését végezzük, ami a referencia-adatbázisban található személyek modelljei és az aktuálisan azonosításra kerülő személy modellje közötti hasonlóság mérését jelenti.

3.3. Érzelemfelismerés

Az érzelmi töltet felismerése viszonylag fiatal ága a beszédfelismerésnek (Sztahó 2014). Napjaink célkitűzése e területen az, hogy 4 ún. alapérzelmet (haragos, örömteli, semleges, bánatos) különítsenek el akusztikai jellemzők alapján gépi osztályozó módszerekkel. A számos felhasználási lehetőségen túl a felhasználó érzelmeinek

követése sokat segíthet a dialógusok dinamikus felépítésében, a beszélő érzelmeire adekvát gépi válasz kiválasztásában, ily módon az ember-gép kommunikáció teljesebbé tételében.

4. Kitekintés

Napjainkban egyre több munkát és anyagi erőforrást összpontosítanak arra, hogy az ember minél természetesebben tudjon érintkezni az őt támogató gépekkel. Ez nem elsősorban a kényelmünket szolgálja, hanem a minket körülvevő nagymennyiségű információ feldolgozásában lehet segítségünkre. Emellett igen nagy szerepe van a valamilyen kommunikációs hátránnyal élők életminőségének javításában is; továbbá azon betegségek diagnosztizálásában is jelentőségük van a beszédtechnológiai alkalmazásoknak, amelyek a beszédben a betegség korai stádiumában produkálnak tüneteket (pl. a gége eltérései hallhatók a zöngeminőségben; az Alzheimer-kór a beszéd időzítésében korán tetten érhető). A mesterséges intelligencia kifejlesztése közben a humán gondolkodás sajátosságait is mélyebben megismerhetjük.

Irodalom

Belin, Pascal – Fecteau, Shirley – Bédard, Catherine 2004. Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences* 8/3. 129–135.

Bimbot, Frédéric – Bonastre, Jean-François – Fredouille, Corinne – Gravier, Guillaume – Chagnolleau, Magrin Ivan – Meignier, Sylvain – Merlin, Teva – Garcia, Ortega Javier – Petrovska-Delacrétaz, Dijana – Reynolds, Douglas A. 2004. Tutorial on text-independent speaker verification. In: *Proceedings of EURASIP, Journal on Applied Signal Processing* 4. New York, USA. 430–451.

Bőhm Tamás 2007. Beszélőfelismerés – neurológiai háttér és pszichológiai modellek. *Magyar Pszichológiai Szemle* 62/4. 541–563.

Dommelen van, Wim A. – Moxness, Bente H. 1995. Acoustic parameters in speaker height and weight identification: Sex-specific behaviour. *Language and Speech* 38. 267–287.

Gocsál Ákos 1998. Életkorbecslés a beszélő hangja alapján. *Beszédkutatás 1998.* 122–134.

Gósy Mária 2001. A testalkat és az életkor becslése a beszéd alapján. *Magyar Nyelvőr* 125/4. 478–487.

Gósy Mária 2016. Beszédhangok viselkedése a spontán beszédben. In Bóna Judit (szerk.): *Fonetikai olvasókönyv.* ELTE Fonetikai Tanszék, Budapest, 19–31. www.fonetikaitanszek.hu.

Lass, Norman J. – Hughes, Karen R. – Bowyer, Melanie D. – Waters, Lucille T. – Bourne, Victoria T. 1976. Speaker sex identification from voiced, whispered and filtered isolated vowels. *Journal of the Acoustical Society of America* 59. 675–678.

Németh Géza – Olaszy Gábor szerk. 2010. *A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek.* Akadémiai Kiadó, Budapest.

Nikléczy Péter 2001. A műszeres személyazonosítás lehetőségei rövid időtartamú beszédminták alapján. *Beszédkutatás 2000.* 154–172.

Nikléczy Péter – Gósy Mária 2008. A személyazonosítás lehetősége a beszédanyag időtartamának függvényében. *Beszédkutatás 2008.* 172–181.
Olaszy Gábor 1999. Beszédadatbázisok készítése gépi beszéd-előállításhoz. *Beszédkutatás* '99. 68–89.
Olaszy Gábor 2010a. Beszédből készített elembázisok beszédszintézishez. In Németh Géza – Olaszy Gábor (szerk.): *A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek.* Akadémiai Kiadó, Budapest. 283–310.
Olaszy Gábor 2010b. Automatikus szövegfelolvasás. In Németh Géza – Olaszy Gábor (szerk.): *A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek.* Akadémiai Kiadó, Budapest. 429–445.
Pandur Balázs 2011. *Üzenetkezelő rendszer vak és látássérült felhasználók részére mobil eszközökön.* TDK-dolgozat. BME, Budapest.
Ptacek, Paul H. – Sander, Eric K. 1966. Age recognition from voice. *Journal of Speech and Hearing Research* 9/2. 273–277.
Scherer, Klaus R. – Banse, Rainer – Wallbott, Harald 2001. Emotional inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology* 32/1. 76–92.
Sztahó Dávid 2014. Automatikus érzelem-felismerés akusztikai paraméterek alapján. PhD értekezés. BME TMIT, Budapest.
Varjasi Gergely 2016. Beszédadatbázisok. In Bóna Judit (szerk.): *Fonetikai olvasókönyv.* ELTE Fonetikai Tanszék, Budapest, 233–244. www.fonetikaitanszek.hu.

Kérdések, feladatok

1. Gyűjtsön az irodalomból, a filmtörténetből olyan alkotásokat, amelyekben beszélő számítógépek szerepelnek. Hogyan kommunikál ezekben az esetekben a gép és az ember? Milyen hasznuk van ezeknek a gépeknek?
2. Képzelje el, hogy egy vállalat Önt kéri fel, hogy a számítógép vezérelt telefonközpontjához válasszon beszélőt! Milyen szempontokat venne figyelembe, amikor kiválasztja az illetőt?
3. Folytassa a *Jó napot!* hangsor bemondatásához szükséges diád elemek felsorolását! Ugyanezt a hangsort milyen triád elemekből lehetne létrehozni?
4. Hallgassa meg a napi időjárás-előrejelzést a metnet.hu oldalán! Mely pontokon tér el a bemondás a természetes magyar beszédtől, hol hangzik furcsán? Mi lehet ennek az oka?
5. Tesztelje az okostelefonján vagy személyi számítógépén a magyarra is létező beszédfelismerőt! Milyen minőségben működik eltérő háttérzajban, illetve különböző beszédtempó, hangerő esetén?