A MAGYAR BESZÉD AUTOMATIKUS SZINTÉZISÉNEK ELSŐ LÉPCSŐJE

Kiss Gábor–Olaszy Gábor

A magyar beszéd mesterséges előállítására irányuló törekvések új keletűek. Magyarországon csak az utóbbi években alakultak ki azok a feltételek, amelyek lehetővé tették, hogy a magyar beszéd mesterséges előállítására irányuló munkák megkezdődjenek. Köztudott, hogy a szintetizált beszéd a tudományos kutatás, a kommunikációs szolgáltatások, az anyanyelvi oktatás, a számítógépipar, valamint a társadalom számos területén felhasználható. Az igények egyre jobban sürgetik, hogy a magyar beszédre vonatkozóan is megszülessenek azok a nyelvészeti-technikai kutatási eredmények, amelyek alapján számítógépeket, mikroprocesszoros berendezéseket meg lehet „tanítani" magyarul beszélni. Az ilyen kutatások mind nyelvészeti-fonetikai, mind pedig technikai szempontból igen bonyolultak, és speciális képzettségű szakembereket igényelnek. Magyarországon többek között a Posta Kísérleti Intézetben (PKI) és az MTA Nyelvtudományi Intézetének fonetikai osztályán folynak kísérletek a magyar beszéd mesterséges előállítására. A PKI munkatársa, Takács György már 1978-ban kifejlesztett egy olyan számítógépprogramot és hardware-t, amellyel úgynevezett félszintetizált beszédet lehet használni a megváltozott telefonszámok automatikus közlésére. Ez a rendszer számokat tud kimondani 0-tól 9-ig és néhány mondatot. A félszintetizált beszéd lényege abban áll, hogy a számokat, valamint az információ szolgáltatásához szükséges mondatokat természetes emberi beszédből állítják össze, számítógépes mintavételezéssel rögzített szótagok vagy szórészek összekapcsolásával.

A magyar beszéd szabályokon alapuló szintéziséhez az MTA Nyelvtudományi Intézete fonetikai osztályán folyó alapkutatások eredményei jelentik az első lépéseket. Munkánk során olyan számítógépprogramot (FOPRO) hoztunk létre (vö. Kiss–Olaszy MFF 10. 1982), amelynek segítségével el tudtuk végezni azokat a szintetizálási munkákat, amelyek a magyar beszédhangok akusztikai szerkezetének teljes feltárásához kellettek. A munka során létrehoztunk egy olyan adatbázist (vö. Olaszy MFF 8. 1981), amely tartalmazza a magyar beszéd legfőbb alkotóelemeit képező # CV , VCV és VC # hangkapcsolódások és néhány CC-kapcsolat akusztikai szerkezetének létrehozásához szükséges adatokat, így alap lehet a későbbi, automatikus beszédszintézis hardware és software rendszerének kidolgozásához. Az automatikus szintézishez azt kell elérni, hogy a számítógép (vagy mikroprocesszor) a megadott szabályok és adatok segítségével a mesterséges beszédet a természetes beszéd tempójában közvetlenül az utasítás megkapása után azonos idejű üzemmódban generálni tudja. Természetesen az ilyen „beszélő" rendszerek vagy gépek elkészítéséhez szükséges szellemi és műszaki ráfordítás mértéke arányban áll a kitűzött feladat szintjével. Általában még a fejlett ipari országokban is – ahol a mesterséges beszédet már számos területen alkalmazzák – leginkább a korlátozott, adott szószámot igénylő feladatoknál használják a számítógépbeszédet. A meghatározott szókészlet előállítása nagyságrendekkel kisebb feladat, mint a tetszőleges szöveget meghangosító, ún. „text to speech" szintézis. Az előbbinél ugyanis csak a megadott szavak előállításához szükséges adatokat és szabályokat kell a számítógép számára kidolgozni, ebből következik, hogy viszonylag kis adatmennyiséggel kell dolgozni.

Az utóbbi megoldásnál a szintetizálni kívánt nyelvre vonatkozó szinte összes szabályt és adatot – a kivételeket is – ki kell dolgozni, ami komoly és hosszú kutatómunkát igényel, nem is beszélve a hatalmas adattömegről, amit a számítógépnek kezelnie és mozgatnia kell a szintéziskor.

Osztályunkon a magyar beszéd automatikus szintézisének első lépcsőjeként az MFF 8-ban ismertetett adatbázis felhasználásával „Számok" néven készítettünk automatikus szintetizáló programot.

A szintézishez PDP 11/34-es számítógépet és OVE III típusú beszédszintetizátort használtunk. A program és a számítógép segítségével bármilyen számot ki tudunk mondatni a szintetizátorral 1-től 1 billióig. A számítógép értelmezi az alapvető számtani műveleteket és az azokkal felépített képleteket is, kívánságra azokat ki is mondja, közben kiszámítja a képlet végeredményét,és azt hangosan közli. Például a 3617-es szám leütése után a számítógép a következőt mondja: *háromezer-hatszáz tizenhét*. A 2x(49+97)/4 = ? karaktersorozat beütése vagy megadása után a következőt halljuk: *kettőször – negyvenkilenc – plusz – kilencvenhét – per – négy – egyenlő – hetvennégy* vagy a 864–523 jelsorozatra kimondja a telefonszámot. Az alábbiakban ismertetjük a „Számok" program előkészítésének fázisait, rendszerét és működését.

Az akusztikai építőkockák megtervezése

Ha számítógéppel minden számot meg akarunk szólaltatni 1-től 1 billióig, akkor első lépésben egy olyan mátrixot kell készítenünk, amelyik tartalmazza a számok felépítésének összes variációs lehetőségeit. A műveleti jelek megszólaltatásához a mátrixba be kell építeni az összeadás, kivonás, szorzás, osztás szavait, valamint az *egyenlő* kifejezést is. A mátrix első oszlopában helyezkednek el azok a számok és számrészek, amelyek alapelemként szolgálnak az egy- és többjegyű számok előállításához. Összesen 23 ilyen alap-építőkockát különböztetünk meg. Ezek a következők: *1,2,3,4,5,6,7,8,9,10,-tizen-, 20,huszon-,30,40,50,60,70,80,90,100,1000* és *millió*. Ezek helyes kapcsolódásaiból felépülhet minden szám, amely 1 és 999999999 közé esik. A helyes kapcsolódások szintaktikai szabályai:

<SZ 1> = | egy | kettő | három | négy | öt |
| hat | hét | nyolc | kilenc |

<SZ 11> = | tíz | húsz |

<SZ 12> = | harminc | negyven | ötven | hatvan |
| hetven | nyolcvan | kilencven |

<SZ 111> = | tizen | huszon |

<SZ AZ> = | száz |

<EZER> = | ezer |

<MILLIÓ> = | millió |

<SZ 10> | <SZ 11> |
| <SZ 12> |
| <SZ 12> |
| <SZ 12><SZ 1> |

<SZ 100> = |<SZ 1><SZÁZ> |
|<SZ 1><SZÁZ><SZ 10> |
<SZ 100> = |<SZ 100><EZER> |
|<SZ 100><EZER><SZ 100> |
<SZ 1000000> = |<SZ 100><MILLIÓ> |
|<SZ 100><MILLIÓ><SZ 1000> |
SZÁM = |<SZ 10000.000> |
|<SZ 1000> |
|<SZ 100> |
|<SZ 10> |
|<SZ 1> |

Az alapelemeket 1-től 23-ig számozott azonosítóval jelöljük. A mátrix első sorában ugyanezek az elemek és a műveleti jelek foglalnak helyet. Ezek képviselik a kettő vagy többelemű számsor következő elemét (1. ábra).

A mátrix első oszlopában lévő alapelemek sorainak és az első sorban lévő „következő" számok, illetve műveleti jelek oszlopainak találkozásánál találhatók az egyes számkombinációk létrehozásához szükséges csatoló elemek. Ezeket is azonosítóval láttuk el 24-től felfelé. Természetesen minden vízszintes sorbeli elem nem találkozhat minden függőleges oszlopbelivel. Például az *egy* után csak a *száz, ezer* vagy *millió* állhat. Az *1000* után viszont az *1,2,3,4,5,6,7,8,9,10, tizen,huszon, 20,30,40,50,60,70,80,90-*nel kezdődő elemek adnak értelmes számot. A mátrix természetesen tartalmazza a hangsor eleji és hangsor végi szünetet is mint építőelemet.

Tulajdonképpen a mátrix első oszlopában lévő 23 számelem felhasználásával már minden számot össze lehetne állítani. Ehhez csak a megszólaltatni kívánt számhoz tartozó elemeket kellene a programnak egymás után kapcsolnia. Például a 717-es számot a *hét – száz – tizen – hét* elemek összekapcsolásával kapnánk meg. Ezt azonban csak az írás szintjén tehetjük meg így. A beszéd szintjén, hogy a számok hangzását a leghűbben megvalósítsuk, figyelembe kell venni a hasonulásokat, a kapcsolódási szabályokat, valamint azokat az időszerkezeti szabályokat is, amelyek a számhangsorok képzésére érvényesek. Vizsgálataink során kimutattuk, hogy a számhangsorok kimondásakor ugyanazon elemek időtartama egy számon belül változó, attól függően, hogy milyen számelem környezetében van. Például a *huszonkettő* és a *huszonnyolc* számok esetében a második szótagokban lévő [o] hangok időtartama nem ugyanaz. Továbbá a mérések kimutatták, hogy egyes számelemek magánhangzóinak időtartama más a hangsor eleji helyzetben és más hangsor záró helyzetben. Például a *húszezer-kettőszázhúsz* számban az első *húsz* időtartama rövidebb, mint a másodiké.

Szintetizálási kísérleteink igazolták, hogy az ilyen és ehhez hasonló jelenségek figyelembevétele lényegesen javítja a gépi beszéd minőségét.

Az akusztikai építőkockákba dallamot nem terveztünk, tehát a kimondott számok és műveletek egyenlő alaphangmagasságon szólalnak meg. Az előbbiek alapján kialakított mátrixból kiolvasható, hogy a számok és a műveletek meghangosításához maximum 340 építőelemre van szükség. Mivel az egyes csatlakozási formák és hasonulások nemcsak egy számkombináció előállításánál fordulnak elő, hanem többnél is; a tényleges elemszám a 340-nél lényegesen kevesebb. Például a [t] és [s] hang találkozásakor kialakuló

A következő szám

		1	2	3	4	5	6	7	8	9	10	20	30	40	50	60	70	80	90	10-en	20-on	100	10^3	10^6	szó vége
Számok	Azonosítók																								
1	1																					25	26	27	28
2	2																					30	31	32	33
3	3																					35	36	37	38
4	4																					40	41	42	43
5	5																					45	46	47	48
6	6																					45	46	47	48
7	7																					45	46	47	48
8	8																					45	51	52	53
9	9																					45	51	52	53
10	10																						54	55	56
20	11																						57	58	59
30	12	24	29	34	39	44	34	49	50	29													51	52	53
40	13	60	61	63	64	65	63	63	66	61													60	67	68
50	14	60	61	63	64	65	63	63	66	61													60	67	68
60	15	69	70	71	72	73	71	71	74	70													75	69	76
70	16	60	61	63	64	65	63	63	66	61													67	60	68
80	17	69	70	71	72	73	71	71	74	70													75	69	76
90	18	60	61	63	64	65	63	63	66	61													67	60	68
10-en	19	60	61	63	64	65	63	63	66	61															
20-on	20	77	78	79	80	81	79	79	66	78															
100	21	83	84	85	86	87	85	88	89	84	84	85	85	86	87	85	88	89	84	84	85		90	83	91
10^3	22	24	29	34	39	44	34	49	50	29	29	34	34	39	44	34	49	50	29	29	34				29
10^6	23	24	29	34	39	44	34	49	50	29	29	34	34	39	44	34	49	50	29	29	34				29
szó eleje		24	29	34	39	44	34	49	50	29	29	34	34	39	44	34	49	50	29	29	34				

1. ábra
A lehetséges kapcsolódásokat realizáló mátrix

hasonulás ugyanaz az *ötszáz, hatszáz* és a *hétszáz* szavakban. Ebből következik, hogy a mátrix egyes belső csatoló elemei egyformák lesznek. Így kialakul az a végleges építőelemszám – 119 –, amellyel a feladatot meg tudjuk oldani (vö. 1. ábra).

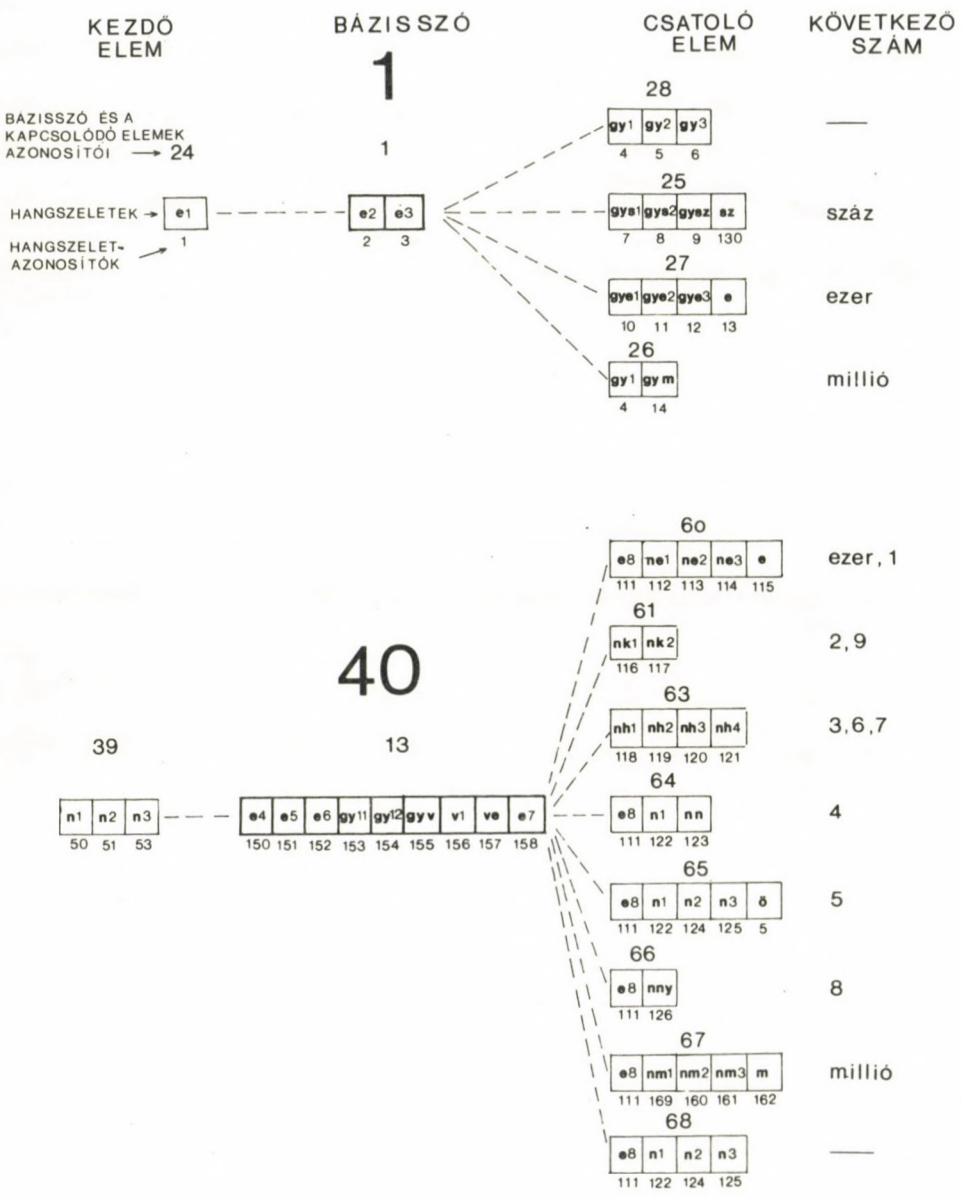
Az adatbázis elkészítése

Az automatikus szintézishez a számítógépprogramnak szüksége van az előbb említett mátrix elemeit realizáló hangszeletek adataira. A hangszeleteket az MFF 8-ban ismertetett adatbázis alapján állítottuk össze. A mátrixot képviselő adatbázist úgy alakítottuk ki, hogy a SZÁMOK programnak a lehető legkevesebb lépésre legyen szüksége a szintézishez. Ez lényeges szempont, hiszen az azonos idejű üzemmód megköveteli, hogy a műveletek számát a legminimálisabbra csökkentsük. Csak így lehet biztosítani, hogy a számítógép el tudja végezni a szám értelmezését, a helyiértékek megállapítását, a szükséges bázisszavak és a csatoló elemek kiválasztását, összeállítását, összekapcsolását a számhoz, valamint a műveleti jelek előhívását és az azokkal járó ragok hozzákapcsolását. Az optimális adatbázis kialakításához bevezettük az ún. bázisszó-építőkockát és a kapcsolódó csatoló elemet. A 23 alapelem mindegyikére meghatároztuk azt a hangsorrészt, amelyik minden előfordulási variációban ugyanaz, hangzása állandó. Ez a bázisszó. A bázisszóhoz kapcsolódik a kezdő elem és a változó akusztikai részeket tartalmazó csatoló elem. Így minden számot a bázisszóval és az eléje és utána kapcsolt elemmel valósítunk meg. A bázisszó mindig állandó, a kapcsolódó elem attól függően változik, hogy a szám után mi következik. Ha a meghangosítandó szám több számjegyű, akkor azt a megfelelő bázisszavak és kapcsolódó elemek sorozatából állítjuk elő. Példaként az 1-es és a 40-es szám bázisszavát és a számhoz kapcsolódható elemeket a 2. ábrán mutatjuk be.

A 2. ábra szerinti felosztást elvégeztük mind a 23 alapelemre, így megkaptuk a mátrix tényleges elemeit. Az első oszlopba kerül tehát a 23 bázisszó és a szóeleji szünet. A többi oszlopot pedig a megfelelő kezdő és csatoló elemmel töltjük meg. Ezen elemek azonosítóiból leolvasható, hogy mely számkombinációknál használjuk ugyanazokat a kezdő vagy csatoló elemeket. Ezek azonosítója ugyanis egyforma.

A bázisszavakat és a kapcsolódó elemeket hangszeletekből építettük fel. A hangszeleteket szintén 1-től kezdődően azonosítóval láttuk el (vö. a 2. ábrán). Természetesen ezeket az azonosítókat a program más helyen tárolja, mint a bázisszeletekét és a kapcsolódó elemekét. A hangszeleteknél is előfordult az, hogy ugyanazon hangszeletet fel lehetett használni különböző bázisszavak vagy kapcsoló elemek felépítésénél. A hangszeletek számának ilyen optimalizálásával a memóriaigényt sikerült csökkenteni. A bázisszavak,a kezdő és kapcsoló elemek és a műveleti jelek felépítéséhez összesen 209 hangszeletet használtunk fel.

A számítási műveleteket, vagyis képleteket a KEPLET nevű számítógépprogramunk hangosítja meg. A programot legegyszerűbben úgy mutathatjuk be, ha leírjuk működését, amely a következő lépcsőkben valósul meg egy adott képlet meghangosítása során.



2. ábra
Példa a szintetizált számok felépítésére

1. Kezdeti betöltés, a floppy disken tárolt adatbázisból a szükséges szeletek beolvasása a központi memóriába.
2. A „~" (tilde) karakter kiírása, evvel jelzi a számítógép, hogy készen áll a meghangosítandó képlet leolvasására.
3. A felhasználó által a képlet beolvasása.
4. A képletsorozat szintaktikai elemzése: a számok, műveleti jelek elkülönítése.
5. A képlet értékének kiszámítása, lengyel formára hozás segítségével, az eredménynek a szintaktikai sorozathoz való csatolása.
6. A szintaktikai egységek sorozata alapján a megszólaltatandó szeletek, hangrészek összeválogatása.
7. A kiválasztott hangrészek egymás utáni lejátszása, meghangosítása.
8. A feladat 2-es ponttól való folytatása.

*

A KEPLET program tapasztalatai alapján 1981–82-ben továbbléptünk, és elkészítettük az UNIVOICE szintetizáló rendszert, amelynek segítségével nemcsak számokat, de bármilyen magyar szöveget real time meghangosíthatunk. A program 365 hangszelet segítségével minden magyar V, VV, CV, VCV, VC és CC hangsorépítő elemet elő tud állítani, értelmezi a leggyakoribb hasonulásokat, valamint a megadott hangsorra kijelentő és felszólító dallamformát is rá tud ültetni. Ez a magyar nyelvű folyamatos text-to-speech szintézis első lépcsője. Az UNIVOICE rendszert a 8. Budapesti Akusztikai Kollokviumon mutattuk be. A rendszerről az MFF egyik későbbi számában számolunk be részletesen.

Irodalom

KISS Gábor–OLASZY Gábor: Interaktív beszédszintetizáló rendszer számítógéppel és OVE III szintetizátorral. MFF 10. 1982, 21–46.

OLASZY Gábor: Hangsorok számítógépes formánsszintézisének előkészítése. MFF 8. 1981, 147–60.

THE FIRST STEP IN THE AUTOMATIC SYNTHESIS OF HUNGARIAN

Gábor Kiss–Gábor Olaszy

This paper gives an account of the first Hungarian real time speech synthesizing program. The results of the research done at the Phonetics Department of the Institute of Linguistics of the Hungarian Academy of Sciences constitute the acoustic basis of the synthesis. A PDP 11/34 computer and an OVE III synthesizer were used in the synthesis. The research was focussed on the formant synthesis of the Hungarian speech by rule. Our main goal in making the program was to try out the data of the data basis gathered as a result of the research done so far. The program makes numbers audible, from 1 to 999 million. When we press e.g. the key 3617, on the keyboard of the typewriter and then press „return", the computer immediately starts to feed the data into the synthesizer and the number becomes audible. There are 23 basic elements in the data basis from which any number can be made up. These are the Hungarian equivalents of (words or word-parts) *1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90. ...-teen, twenty-..., 100, 1000 and million.*