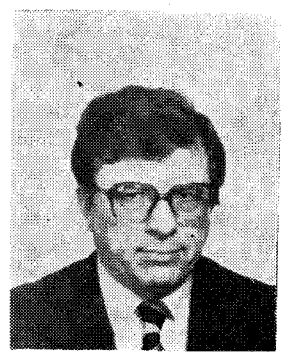
Szolgáltatásbővítés gépi beszédfeldolgozással*

DR. GORDOS GÉZA
Rudapesti Műszaki Egyetem
Híradástechnikai Elektronika Intézet



ÖSSZEFOGLALÁS

A beszéddel kapcsolatos emberi funkciók — a beszélés, megértés és beszélőfelismerés — gépi megvalósításának elvi alapjait a telefónia vetette meg az érthetőségvizsgálatok, a PCM és a vokóder korai megalkotásával. Elsősorban a mikroelektronika megvalósító erejétől támogatva a gépi beszédfeldolgozás szárba szökkent és most viszonoz: új szolgáltatásokat kínál a telefóniának és az egész távközlésnek. E szolgáltatások egyik csoportja a hagyományos szolgáltatásokon belül jelentkezik. Ilyen az előfizetőt szóban tájékoztató vagy útbaigazító telefonközpont. A szolgáltatások másik csoportja merőben új, amit a beszédválaszú, vagy a beszéddel kérdezhető adatbázisok, helyfoglaló rendszerek példáznak.
A dolgozat a szolgáltatások áttekintése után a legelterjedtebb beszédfeldolgozási ág: a beszédszintézis általános módszereit tárgyalja. Ezután néhány eredeti hazai beszédszintézis technológiát ismertet, majd ezek eredményeit összehasonlítja a világszínvonallal. Végül bemutatásra kerül az, hogy a beszédszintézis módszerei miként vezethetnek beszédmegértéshez.

DR. GORDOS GÉZA

1937-ben született, 1960-ban villamosmérnöki, 1966-ban egyetemi doktori, 1977-ben kandidátusi oklevelet szerzett. Fő munkahelye 1960-tól a BME Híradástechnikai Elektronika Intézete, ill. annak jogelődje, ahol jelenleg az átvitel- és rendszertechnika osztályt vezeti. 1964 és 1972 között a Posta Kísérleti Intézetben, 1972-ben UNESCO-szakértőként Görögországban, 1974/75-ben vendégprofesszorként Angliában dolgozott. Fő érdeklődési területe a fém- és fényvezetős digitális átvitel, adatátvitel, valamint a gépi beszédszintézis és beszédfelismerés.

1. Bevezetés

A távközlés fejlődése két alapvető irányban zajlik. Az egyik a már meglevő szolgálatok tökéletesítését, olcsóbbá és megbízhatóbbá tételét szolgálja. Ebbe az irányba olyan hatalmas jelentőségű fejlesztések esnek, mint a tárolt programvezérlés vagy a digitális eljárások elterjedése a kapcsolástechnikában, a fénytávközlés az átviteltechnikában, vagy a szolgálatok integrálódása. A távközlés fejlődésének másik iránya az előfizetőnek nyújtott szolgáltatások bővítése. Ez megjelenhet teljesen új szolgálatok létrejöttében — amit az adatátviteli szolgálat bevezetése példáz. De megjelenhet a már meglevő szolgálatok új vagy fejlettebb szolgáltatásainak formájában is. A távbeszélő szolgálat erre a hívásátirányítás szolgáltatásával, az adatátviteli szolgálat pedig az operátori kísérő hang átvitelének szolgáltatásával mutat példát.

A gépi beszédfeldolgozás leglátványosabban egy sereg új szolgáltatásra és emellett sok régebbi szolgáltatás minőségjavítására kínál lehetőséget. Kevésbé látványos, de nagy jelentőségű az, ahogy a gépi beszédfeldolgozás a már ma is működő szolgáltatások hatékonyságát növelni tudja. Elöljáróban le kell szögezni, hogy a lehetőségek kiaknázása világszerte rohamléptekkel halad előre, és szerény eredmények már hazánkban is mutatkoznak.

2. A gépi beszédfeldolgozás fogalma

A gépi beszédfeldolgozáson első közelítésben az emberi beszédfunkciók mesterséges megvalósítását, illetve azok gépi úton történő utánzását értjük. A fogalom pontosabb értelmezéséhez a természetes beszédláncon [1] keresztül juthatunk el. A természetes beszédlánc az emberi beszélőből, a beszélő által keltett levegőrezgéseket átvivő akusztikus térből s az emberi felfogóból áll. Ezen természetes beszédlánc egy vagy több elemének mesterséges megvalósítását nevezzük gépi beszédfeldolgozásnak.

A gépi beszédfeldolgozás nem új keletű. Első hiteles [2, 3] és sikeres megjelenése Kempelen Farkas nevéhez fűződik, aki 1791-ben egy erősen korlátozott képességű „beszélő gép" megalkotásáról számolt be. A társadalomban széles körben először a természetes beszédlánc második elemének — a szájtól fülig terjedő akusztikus térnek — a mesterséges megvalósítása terjedt el, mégpedig egyrészt a térbeli távolságot legyőző távbeszélő átvitel és rádiózás, másrészt az időbeli távolságot legyőző hangrögzítés formájában. Igaz, hogy e technikák a beszédet minden feldolgozási ponton analóg módon kezelték, amit — e tekintetben értett egyszerűsége miatt — ma nem tekintünk beszédfeldolgozásnak, a ma is fajsúlyosnak ítélt feladatok megoldásának néhány sarkkövét mégis e technikák munkálták ki. A korai telefónia (Fletcher, [4]) munkálta ki az objektív paraméterek (sávszélesség, jel-zaj viszony stb.) és a szubjektív paraméterek (érthetőség, hanghűség) közötti kapcsolatokat, és vetette ezzel meg a percepciós vizsgálatok máig is legfontosabb alapjait.

Jelentőségében ezzel összemérhető Reeves pulzus—kód—modulációs szabadalma [5]. Ez volt az első lépés a híradás- és számítástechnika eljárásait és módszereit ötvöző digitális jelfeldolgozás felé. Az elsőt újabbak követték: a vocoder (Dudley, [6]) és a beszédszüneteket detektáló és kihasználó TASI rendszer már a magasabb emberi beszédfunkciókat valósítják meg elektronikus eszközökkel, és így a mai modern, digitális elektronikán alapuló gépi beszédfeldolgozás közvetlen előfutárainak tekinthetők.

* Elhangzott a Magyar Tudományos Akadémia 1984. november 1-i tudományos ülésszakán.

3. A gépi beszédfeldolgozás osztályozása

Ma a gépi beszédfeldolgozás igen sok ágra bomlik, és a gyors fejlődés az egyértelmű osztályozást nehézzé teszi. Mindazonáltal néhány ág értelmezése egységesnek tekinthető.

3.1. *Beszédtömörítésen* olyan kódolást és dekódolást értünk, amely hatékonyabb a hagyományos PCM-nél, tehát 1 sec-nyi aktív beszéd átvitelét, illetve rögzítését 64 kbit-nél kevesebb ráfordítással oldja meg. Ezt az irányt ma egyebek között a CCITT is erősen kutatja.

3.2. A *beszéddetekció* a beszéd jelenlétét indikálja. A TASI rendszer, ill. az INTELSAT ezt az átviteli utak jobb kihasználására a SPADE és INTELCSAT rendszer pedig az inaktív csatorna adójának kikapcsolására használja fel. Igen nehéz a zajos ipari környezetben működő beszéd jelenlétének detektálása [8], amire beszédfelismerő rendszerek automatikus indításánál, illetve kikapcsolásánál van egyebek között szükség.

3.3. *A beszédszintézis* fogalma önmagát magyarázza. Ma ez a gépi beszédfeldolgozás legszélesebb körben alkalmazott ága, ezért az alábbiakban ennek fogjuk a legtöbb figyelmet szentelni.

3.4. *Beszédmegértés* az elhangzott beszéd jelentéstartalmának gépi felfogása. A felfogás eredménye első fokon a jelentéstartalomnak megfelelő akció: például egy parancs végrehajtása. Egy következő fokozat — és valójában az elsőnek nyelvészeti jellegű továbbfejlesztése — a jelentéstartalom betűképének helyesírás szerinti megjelenítése. A beszédmegértés (korábban beszédfelismerésnek nevezték) a beszédszintézisnél lényegesen bonyolultabb feladat. Bonyolultsága végső soron abból fakad, hogy az emberi beszédmegértés fiziológiai és idegi folyamatai nemcsak hogy kevésbé ismertek a beszédképzés folyamatainál, de lényegében sok ponton még feltáratlanok. Megoldás mégis van, s ez értelmezésünk szerint [9] a beszédszintézis felől, mégpedig törvényszerűen onnan bontakozik ki. Nyilvánvaló ugyanis, hogy a beszélő szervek folyamatait a mondanivaló határozza meg, s a beszédmegértés célja éppen e mondanivaló megállapítása. Ha tehát a hanghullámból sikerül visszakövetkeztetni a beszélő szervek folyamataira — ami a beszédszintézis inverz feladata — akkor közelebb jutottunk magához a mondanivalóhoz. A cikk végén vázoljuk majd ezen elv egyik megvalósítását. Annak dacára, hogy a beszédmegértés még messze van a tökéletestől, sikeres fejlődés két irányban is tapasztalható. Az egyik irány kevés (30...60) izoláltan kiejtett szót bárkinek a kiejtésében nagy valószínűséggel felismerő eszközökhöz vezetett az ezer dollárok nagyságrendjébe eső áron. A másik irány készüléke igen sok (mintegy 1000) szót ismer fel folytonos kiejtésben is, ám a szöveget csak néhány (2...5) személytől fogadja el, akikhez a gépet egy tanulási folyamattal „hozzá kell szoktatni". Az eszközök ára a tízezer dollárok nagyságrendjében van. Az 1. ábrán bemutatott diagramunk talán jól érzékelteti a jelen helyzetet. Több helyen kipróbáltak már mesterséges gépírókat is, s ha azok témákra szakosodnak (szókincs!), akkor az eredmények biztatóak.

3.5. Nemcsak a kriminalisztika, hanem egyre inkább a pénzügyi tranzakciókat, információ szolgáltatásokat stb. kísérő személyazonosítás is jó hasznát veszi a hang alapján történő gépi *beszélő személy azonosításának* és *felismerésének*. Az eljárás automatikus — szemben pl. az ujjlenyomat-vizsgálattal, — és megbízhatósága lényegesen felülmúlja az aláírás alapján történő azonosítást, akár ember, akár gép végzi az utóbbit.

3.6. A digitális beszédfeldolgozás kis méretű készülékekkel gyakorlatilag megfejthetetlen *beszédtitkosítást* tesz lehetővé.

3.7. Itt csak megemlítjük, hogy a gépi beszédfeldolgozást sikerrel alkalmazzák *beszéd-manipulálásra*, igen sok *orvosdiagnosztikai* célra és a *beszédkészség javítására* stb.

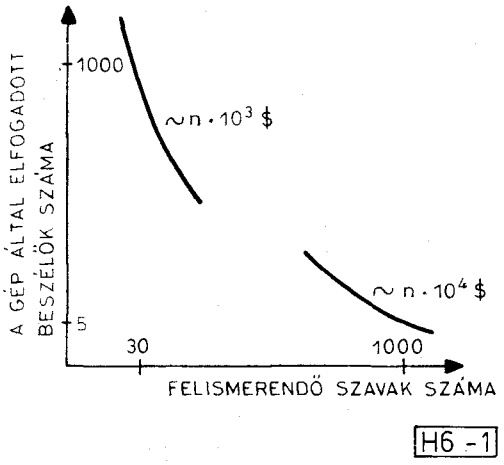
4. Gépi beszédfeldolgozás a távközlésben

A gépi beszédfeldolgozás mai lehetőségeinek fenti felsorolása után a távközlési alkalmazások szinte magától értetődőek.

A távbeszélő-szolgálat hagyományos szolgáltatása a speciális hívószámokon elérhető tájékoztatás (pontos idő, időjárás stb.), illetve mese. Ha itt a magnetofon szerepét beszédszintetizátor veszi át, a szolgáltatás megbízhatósága és a szöveg szerkeszthetősége óriásit javul.

A kis méretek, a mikroelektronizáltság, a szöveg könnyű — akár központi processzorból, automatikusan történő — szerkeszthetősége a beszédszintetizátorokat kiválóan alkalmassá teszik arra, hogy távbeszélő központok előfizetőiket élőszóban tájékoztassák. Az előfizető és a távbeszélő-szolgálat közötti kommunikáció leglényegesebb gátja ma az, hogy az előfizetők tömegei nem tudják helyesen értelmezni a túl sokféle jelzőhangot. A manuális üzemről automatára történő áttérés egyetlen hátrányán — az elszemélytelenő előfizető/szolgálat kapcsolaton — a beszédszintézis már ma, a beszédmegértés pedig a közeljövőben segíthet.

A számítástechnika és távközlés szimbiózisának egyik legnagyobb jelentőségű fejleménye az, hogy nagy tömegek is hozzáférhetnek adat- és információs bázisokhoz. A tömeges elterjedésnek azonban feltétele



1. ábra. A gépi beszédmegértés tipikus teljesítőképessége 1984-ben. $1 \leqq n \leqq 10$

az, hogy az előfizetői terminál valóban olcsó legyen. A terminál input és output funkciókkal rendelkezik, melyek közül az output (képernyő, nyomtató stb.) a drágább. Ezt a drágább funkciót a beszédszintézis meg tudja takarítani, amennyiben az adatbázis a hozzá érkezett kérdésre a választ beszédszintézis útján a távbeszélő vonalon keresztül adja meg, és azt az előfizető a kézibeszélő hallgatóján keresztül veszi. A műszaki lehetőségek ennek a beszédválaszú szolgáltatásnak a bevezetéséhez teljességgel adottak. Az előfizetői terminál input szerepét itt vagy maga a választómű (számtárcsa, billentyűzet), vagy egy néhány ezer forintos klaviatúra szolgálhatja.

Ahhoz, hogy a távbeszélő készülék mindenfajta kiegészítés nélkül legyen alkalmas információs bázisok interaktív lekérdezésére, az információs bázist ki kell egészíteni beszédmegértő egységgel. Tematikában korlátozott információs bázisok (pl. helyjegyfoglalás vagy műsortudakozódás) esetén ez már ma elérhető.

Több jel mutat arra, hogy a beszéd titkosításának is opcionális távbeszélő-technikai szolgáltatássá kell majd válnia.

A fentiekben arra mutattunk példát, hogy a gépi beszédfeldolgozás hogyan jelenhet meg az előfizetők felé. A távközlés belső mechanizmusának tökéletesítésében, hatékonyabbá tételében — leginkább az átviteli utak jobb kihasználásában — a gépi beszédfeldolgozás ugyancsak sokrétű alkalmazást nyer.

5. Ember—gép kapcsolatok

Az ember hagyományosan „kezével" kezeli a gépet, és a szemével figyeli és leolvassa a gép közlendőit. Az ember—gép kapcsolatok hagyományos formái között a gépek akusztikus jelzései elhanyagolható jelentőségűek.

A gépi beszédfeldolgozás egészen új távlatokat nyit ezen a területen. A beszéd az ember legtermészetesebb, legkevésbé fárasztó kommunikációs módja. Beszédkapcsolatnál nem kell kéztávolságon belül tartózkodni, nem kell megfelelő „látószögben" elhelyezkedni, és a fül sokkal alkalmasabb a szimultán figyelésre, mint a szem. Mindezekért az ember—gép kapcsolatokban a beszéd egyre fokozódó szerephez jut, és ez alól a világ talán legkomplexebb gépezete, a távközlés sem kivétel.

6. A beszédszintézis általános elvei

A gépi beszédfeldolgozás legfejlettebb, mindenfajta alkalmazásra kész ága a beszédszintézis.

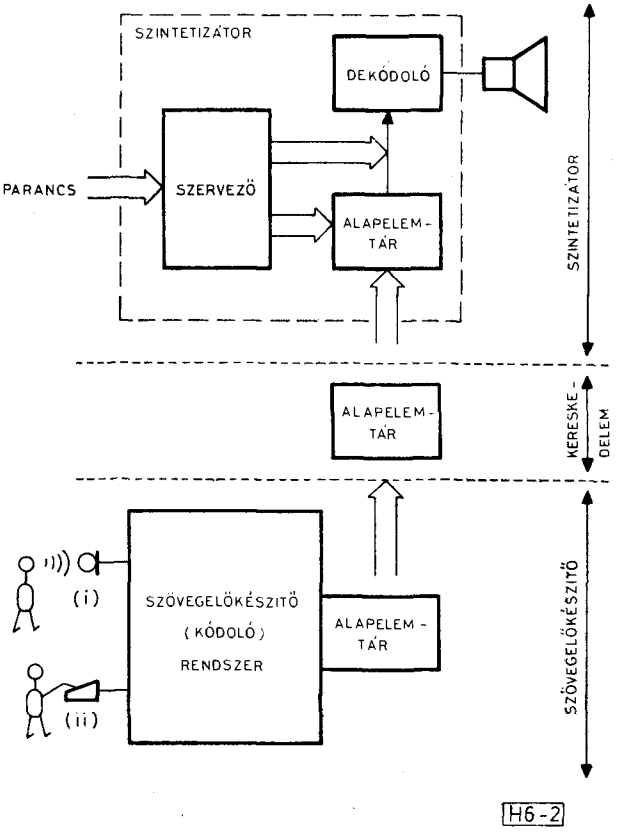
A beszédszintetizáló rendszerek két elemből tevődnek össze a 2. ábra szerint: a *beszédszintetizátorból* és a *szövegelőkészítő rendszerből*. Összefüggésükre később mutatunk rá.

A beszédszintetizátorok általában egy *alapelemtárat*, egy *dekódolót* és egy *szervezőt* tartalmaznak.

Az *alapelemtár* elemei nyelvi egységek. A szintetizátorok bizonyos típusában az alapelemek hangok (pl. „a"), más típusban kettős hangok (diádok, pl. „aj"), megint más típusnál szavak vagy egész kifejezések (pl. „ajtó" vagy „az ajtó záródik").

A *szervező* értelmezi a felhasználói parancsot, és vezérli a szintetizátort. A „szöveget beszéddé alakító" ún. *text-to-speech converter*-eknek a kimondandó közlemény betű/írásjel kódját kell megadni. A szervező — általában egy 20...100 kbyte-os program — ebből megszerkeszti a megfelelő hang- vagy diádalapelemek sorozatát, ezt módosítja az írásjelnek megfelelő szupraszegmentális szerkezettel (hangmagasság-, ritmus- és intenzitásvariációk), és elindítja a kimondást. Ilyen programot ma még csak egy adott nyelvre tudnak készíteni, de ezen a nyelven tetszőleges szöveg megszólaltatható. A szöveg érthetősége és természetessége nagyban függ az éppen használt alapelemek számától és milyenségétől. Ezt az magyarázza, hogy az élő nyelvben egyazon nyelvi egység (pl. az „aj" diád) a szövegkörnyezettől függően végtelen változatossággal jelenik meg. Ma már léteznek olyan text-to-speech szervezők, amelyek 300...600 alapelemet (a legsikeresebb text-to-speech converterekben: diádot) is tudnak kezelni. Hogy pontosan mik legyenek az alapelemek, azt még minden nyelvnél kutatják. A ma elért érthetőség 85%, a természetesség pedig 60% körül van.

Ahol a 100% érthetőség a követelmény, ott az alapelemek szavak vagy rövid kifejezések. Tekintve, hogy minden tár véges, az ilyen szintetizátorok *kötött szótárúak*. Ha azonban arra gondolunk, hogy egy pályaudvari közönségtájékoztatás 50...150 szóval, és az ezek összefűzésével képezhető több ezer értelmes szófüzérrel tökéletesen megoldható, megértjük, hogy miért terjednek az ilyen szintetizátorok. Belátjuk, hogy ez a „kötöttség" nem is nagy ár a



2. ábra. Beszédszintetizáló rendszer
i: Automatikus szövegelőkészítés
ii: Interaktív szövegelőkészítés

100% érthetőségért és 95% feletti természetességért. A kötött szótáras szintetizátoroknál egy-egy közleményre annak szám-, ritkábban közönséges írás szerinti betűkódjával kell hivatkozni. A szervező ezt értelmezi, ilyenekből füzéreket képez, és azon szerkeszt: kihagy, beszúr, kicserél. A szervező program ritkán nagyobb 2 kbyte-nál. A kötött szótáras szintetizátorok alapelemtárai cserélhetők, bővíthetők. Nincs akadálya annak sem, hogy a különböző alapelemek különböző nyelvhez tartozzanak. Sőt, a kötött szótáras beszédszintézis árnyalatokat is ki tud fejezni. (Árnyalaton a kiejtés írásban nem rögzíthető variációit értjük.)

Léteznek olyan szintetizátorok is, amelyek hang és/vagy diád típusú alapelemeket tartalmaznak, és kötött szótárasak. Itt egy közlemény kimondási parancsának megérkezésekor a szervező kikeresi az alapelemek ezen közleményhez előre összeállított sorozatát, és elindítja a kimondást. Ezek a szintetizátorok hangminőség tekintetében alig jobbak a text-to-speech convertereknél.

Még nem esett szó a *dekódolókról*. Mint láttuk, bármilyen típusú is a szintetizátor, szüksége van egy alapelemtárra. Rendkívüli esetektől eltekintve az alapelemek együttesen 30...60 sec-nyi vagy még több beszédet képviselnek. Ennek hagyományos PCM típusú őrzéséhez 2...4 Mbyte-os vagy még nagyobb tároló kellene. Ezt elkerülendő, az alapelemek az általános célú szintetizátorokban tömör formában vannak kódolva. A dekódoló egy olyan aritmetika, amely a tömör, 1000...4000 bit/sec jelfolyamból a fül számára élvezhető 64 000 bit/sec beszédhullámot létrehozza. Megjegyezzük, hogy a kereskedelem — az önmagában életképtelen — dekódolót szokta „beszédszintetizátor"-ként emlegetni.

Fentiek után már nyilvánvaló a *szövegelőkészítő rendszer* szerepe: ez hozza létre a tömör formában kódolt alapelemtárat. Text-to-speech rendszereknél az alapelemtár egyszer s mindenkor adott, itt tehát a szövegelőkészítésnek nincs folytonos szerepe. Kötött szótáras rendszerekben azonban a szótár bővítése csak a szövegelőkészítés közbejöttével végezhető.

7. Tömörítési filozófiák és eljárások

7.1. *A tömörítés minősítése*

Mint láttuk a beszédszintézis kulcskérdése a tömörítés. Ugyanakkor az átviteltechnika is sokat foglalkozik a beszéd 64 kbit/s-nál lassúbb, de kiváló minőséget biztosító átvitelével. A beszéd digitalizálása a digitális hangrögzítésnek is feladata. Egységes képben szemlélhetjük e technikákat, ha a beszéd-digitalizálás jóságát az

$$M = c_1 \cdot H + c_2 \cdot V + c_3 \cdot K \qquad (1)$$

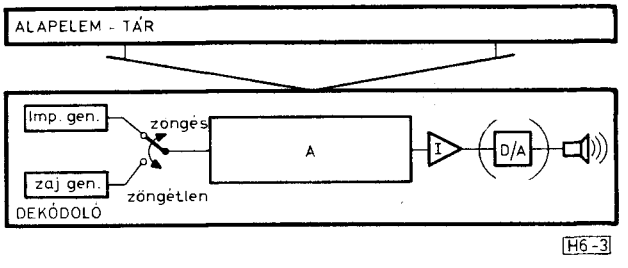
összefüggéssel értelmezett [9] M szám kicsinységével mérjük. Itt H az eredeti és a kódolás—dekódolás folyamata után visszaállított jelek közötti eltérés emberi megítélést tükröző mérőszáma, V az időegységnyi eredeti üzenet megadásához szükséges kódolás utáni bit-szám, K pedig a kódoló és dekódoló bonyolultságának olyan mérőszáma, amely azt is tükrözi, hogy real-time megoldás létezik-e vagy sem. A c_1, c_2 és c_3 súlyozó tényezők az alkalmazástól függenek. A digitális hangrögzítés nagyon szigorú H-nál, míg „engedékeny" V és K területén, bár K-nál megköveteli a real-time működést. A telefónia közepesen szigorú H-nál, de kicsi, olcsó, real-time kódolót és dekódolót követelvén szigorú K-val szemben, s V így 64 kbit/sec körülire adódik. A beszédszintézis valamivel enyhébb H tekintetében, de — szótártól függően — esetenként 3 kbit/sec alatti V-t követel, ezért K-ban kénytelen engedni: a kódoló legtöbbször nem alkalmas real-time működésre, és a real-time dekódoló is több mikroprocesszor bonyolultságú eszköz.

7.2. *PCM alapú és fázisrendező eljárások*

Térjünk ezek után rá a beszédszintézisben alkalmazott tömörítési (kódoló/dekódoló) eljárásokra. Csak igen kis alapelemtárú szintetizátor rendszerek elégedhetnek meg a *PCM*, vagy annak rokonai, a *DM*, *DPCM*, *ADPCM* által nyújtott, a 20 kbit/sec alá jutni nem tudó tömörséggel. Még a *D*igitalker-ben [10] alkalmazott fázisrendezéses eljárás ([1]) által nyersen biztosított 16 kbit/sec sem bizonyul legtöbbször elég tömörnek.

7.3. *Formánsszintézis*

Ezért a 60-as években mind jobban a formánskódolásra terelődött a figyelem. Itt abból indulnak ki, hogy a legtöbb nyelven már elég megbízható adatok állnak rendelkezésre az egyes hangok és hangátmenetek akusztikai szerkezetéről: időtartamáról, relatív spektrumáról és intenzitásáról. Azt is megfigyelték, hogy a hangmagassággal rendelkező hangok spektruma többé (magánhangzók) — kevésbé (felpattanó zöngés hangok) vonalas, míg a többieké folytonos. Előbbieket tehát egy periodikus, utóbbiakat egy zajgenerátor által táplált, a hang spektrumának megfelelően beállított szűrővel létre lehet hozni. Ha másik hangot akarunk kelteni, át kell „hangolnunk" a szűrőt, és a megfelelő gerjesztést kell alkalmazni. A 3. ábrán látható, vezérelhető szűrőt tartalmazó elrendezés tehát egy beszédszintetizátor. A kódolást egy fonetikai ismeretekkel rendelkező személy végzi. Mivel egy-egy hang vagy átmenet 8...200 msec időtartamú, és ezalatt a generátort és szűrőt 10...12 adat á 2...6 bittel határozza meg, a formánsszintézis tömörsége szövegtől függően 600...2000 bit/sec. Ez, és az a tény, hogy 1982-ben egy mikroelektronizált formáns szintetizátor [11] jelent meg a piacon, az



3. ábra. Formánsszintetizátor (az „A" átvivő rendszer formánsszűrő) és LPC/PARCOR szintetizátor (az „A" átvivő rendszer az 5. és 6. ábra szerinti) felépítése.
$\tilde{e}(n)$: normalizált gerjesztés (l. a szöveget)

alkalmazásokat erősen motiválja. Utóbbi magyar nyelvi felhasználására is megtörténtek már az első lépések [12], melyek a közeljövő *MINIVOX* rendszerét ígérik.

Mindeme sikerek mellett látnunk kell, hogy a formánskódolás átlagos nyelvi megfigyelésekre támaszkodik. Ezért a hangzás „személytelen", nem 100%-osan természetes, sőt, csak a legképzettebb kódoló személyek tudják meghaladni a 95%-os érthetőséget.

A formánskódolás klasszikus eljárása tehát nem a kódolandó beszédrész egy konkrét kiejtéséből indul — mint a *PCM*, *DPCM* stb. Ezért mi utóbbiakat „lejegyző", előbbit „generáló" eljárásnak tekintjük. Nyilvánvaló, hogy a formánskódolás minőségén azzal lehetne javítani, ha az is természetes kiejtésből indulna. Ebben az irányban jelentkezett kezdeti eredményekkel 1983-ban az Utrechti Egyetem, s ilyen irányú munkák folynak — a „virtuális formáns" új fogalmára [1] alapozva — a Budapesti Műszaki Egyetemen.

7.3. *LPC és PARCOR eljárások*

Ma a legjobb eredményt egy egészen más elv, a lineáris predikción, illetve parciális korreláción alapuló kódolás/dekódolás adja. Noha az eljárás két alapgondolatát Atal és Hanauer [13], illetve Itakura és Saito [14] csak a 60/70-es évek fordulóján vetette fel, a módszer már célba ért. Ezen az elven alapul a Texas Instruments [15], illetve a Nippon EC mikroelektronizált szintetizátora, s ennek továbbfejlesztésével készült a *BME LIAVOX* beszédszintézis rendszere [16].

Az eljárás megértéséhez tekintsük a beszédet 10...20 msec hosszú szegmensekre bontottnak. Jelölje egy szegmens mintáit $s(0), s(1), \ldots, s(N-1)$. Tegyük fel, hogy minden $s(n)$ minta jól közelíthető az őt megelőző p darab minta $\sum_{i=1}^{p} a_i \cdot s(n-i)$ alakú lineáris kombinációjával. Az a_i ún. *l*ineáris *p*redikciós együtthatókat (=coefficient: *c*; LPC) meg lehet határozni úgy, hogy az

$$e(n)=s(n)=\sum_{i=1}^{p} a_i \cdot s(n-i) \qquad (2)$$

ún. predikciós hiba négyzetösszege, $\sum_{n=0}^{N-1} e^2(n)$, minimális legyen. (Hogy $e(0), e(1), \ldots, e(p)$ is számítható legyen, az $s(-p)=s(-p+1)=\ldots=s(-1)=0$ feltételezéssel élünk, ami [1] szerint sehol sem vezet ellentmondásra.) A 4. ábrából látszik, hogy az $e(n)$ sorozat általában kisebb abszolút értékű elemekből áll, tehát kevesebb bittel adható meg, mint az $s(n)$ sorozat. Mivel az $E=\{e(0), \ldots, e(N-1), a_1, \ldots, a_p\}$ adathalmazból az $S=\{s(0), \ldots, s(N-1)\}$ adathalmaz az $s(n)=e(n)+\sum_{i=1}^{p} a_i s(n-i)$ összefüggéssel pontosan visszaállítható, de E megadásához kevesebb bit kell, mint S megadásához, adattömörítést értünk el. A pontos visszaállítás miatt jogos az eljárást *„hullámforma kódolás"*-nak nevezni.

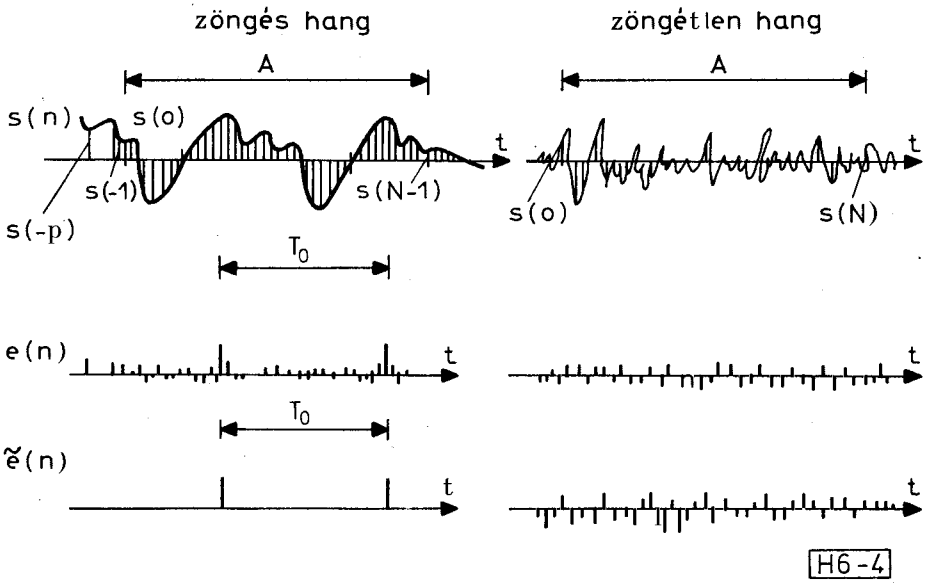
Az így elérhető 3—4-szeres *tömörítés 10—20-szorosra fokozható* az alábbi felismeréssel. Az ábrából látható, hogy zöngés beszédhez tartozó szegmensnél $e(n)$ helyettesíthető olyan — az alapfrekvenciával megegyező periodicitású — impulzussorozattal, amelyben csak egy vagy néhány minta különbözik zérustól.

Zöngétlen szegmenseknél viszont az $e(n)$ egy, a hangtól független zajgenerátor jelének tűnik, melynek csupán a „nagysága" függ az éppen vizsgált szegmens tényleges jelétől. Jelölje $e(n)$ fentiek szerinti közelítését $\tilde{e}(n)$ (l. a 4. ábrát). Atal és társai arra a meglepő felismerésre jutottak, hogy ha a visszaállításhoz $e(n)$ helyett $\tilde{e}(n)$-et használjuk, az

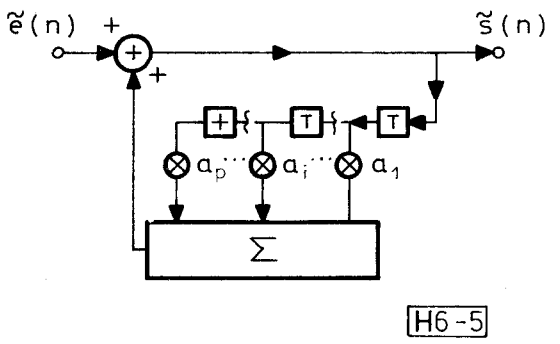
$$\tilde{s}(n)=\tilde{e}(n)+\sum_{i=1}^{p} a_i \cdot \tilde{s}(n-i) \qquad (3)$$

sorozat tökéletes érthetőségű és tökéletes természetességű beszédhez vezet, sőt, legtöbbször $\tilde{s}(n)$-ből még az a személy is felismerhető, akitől az $s(n)$ származik.

A lineáris predikció elvén alapuló tömörítés (Linear Predictiv Coding = *LPC*) dekódolója ezek után ugyanúgy tartalmaz impulzus- és zajgenerátort, mint



4. ábra. A lineáris predikció értelmezése. T_0: alapperiódus (hangmagasság-)idő. A: szegmens

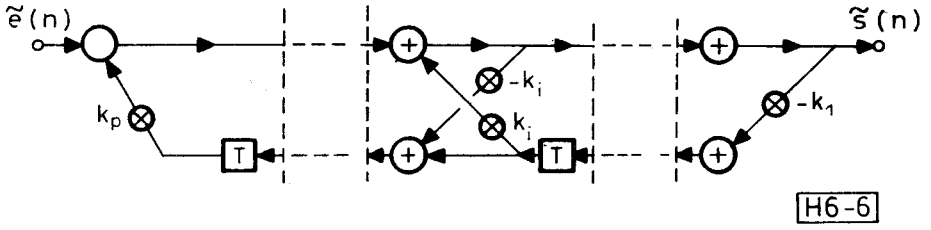


5. ábra. LPC dekódoló

kurzív, másrészt a struktúra stabilitása a $|k_i| \leqq 1$ feltételekkel egyszerűen ellenőrizhető, ami praktikus szempontból igen fontos. A *LIAVOX* rendszer a *PARCOR* struktúrát valósítja meg [16].

8. A beszédmegértés felé...

Fant-tól [17] származik az az ötlet, hogy a beszédképző szerveket (hangszálak, rések, garat-, száj-, orrüreg) a 7. ábra szerint egy dugattyúval és egy azonos hosszúságú (l), de lépcsőzetesen változó keresztmetszetű (A_i) csőszakaszokból álló rendszerrel model-



6. ábra. PARCOR dekódoló

a formánsszintézis dekódolója (3. ábra). A szűrő helyébe azonban itt a (3) egyenletet megvalósító elrendezés lép, melyet önmagában az 5. ábra mutat be. Gyakorlati okokból legtöbbször itt is egységnyi amplitúdójú impulzus, és egységnyi szórású zajgenerátort alkalmaznak (melyek jelét $\tilde{\tilde{e}}(n)$-nel jelöljük). A dinamikaviszonyok helyreállítására a LIAVOX rendszer a $\sum_N s^2(n) = \sum_N [\tilde{s}(n)]^2$ feltételből származó I „gain factor"-t használja jó eredménnyel.

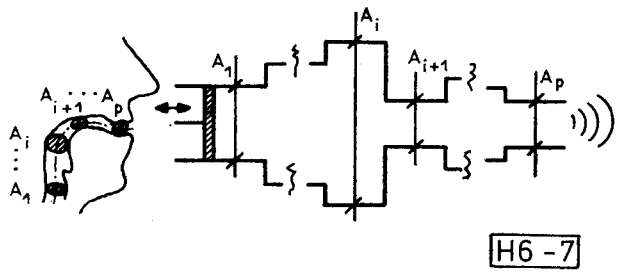
Az *LPC* beszédszintézis tehát az alábbi fázisokból áll:

— szóelőkészítés fázisa: szegmentálás (tipikusan 10...25 msec-nyi részletek, $N=80...260$ mintával), zöngés—zöngétlen jelleg és előbbi esetben a kvázi-periódusidő megállapítása, a hibaminimalizálást megvalósító $a_1, a_2, \ldots, a_p$ ún. *LPC* együtthatók — általában komoly jelfeldolgozást involváló, itt nem részletezett — meghatározása (p értéke tipikusan 8.. 12), „gain factor" meghatározása; azaz szegmensenként összesen $(p+2)$ adat meghatározása és elhelyezése az alapelemtárban;

— a szó kiejtésének fázisa: a 3., 5., ill. 6. ábrán bemutatott *LPC* szintetizátor működtetése úgy, hogy szegmensről szegmensre a vezérlés beállítja a dekódoló paramétereit, melyek azután a szegmens időtartama alatt változatlanok maradnak.

A szóelőkészítés igen magas fokon automatizálható, és mivel nem nyelvészeti szabályokra, hanem a lineáris predikció matematikai elméletére épít, nyelvfüggetlen, sőt, minden olyan hangjelenséget visszaad, amit az ember egyáltalán kelteni tud (nevetés, köhögés stb.). Ezt a magas minőséget a zöngés/zöngétlen döntésben, periódusmeghatározásban és lineáris predikcióban megtestesülő mély elméleti apparátus biztosítja.

Ha az 5. ábrán bemutatott dekódoló helyett a 6. ábrán bemutatott ún. *PARCOR* struktúrát használjuk, két előnyhöz jutunk. Egyrészt a benne szereplő k_i parciális korrelációs együtthatók számítása re-



7. ábra. Az emberi beszédkeltés akusztikus csőmodellje

lezzük. Ezt a modellt $T=2l/v$ (ahol v a hang terjedési sebessége) időközönként vizsgálva, a csőrendszert ugyanaz a diszkrét idejű átviteli függvény írja le, mint a 6. ábra szerinti *PARCOR* szintézis struktúrát, ha a $k_i=(A_i-A_{i+1})/(A_i+A_{i+1})$ megfelelést felismerjük. Ez viszont azt jelenti, hogy ha a szintézisnél ismertetettek szerint meghatározzuk k_i-t, abból a beszélő szervek A_i/A_{i+1} keresztmetszeti viszonyai megállapíthatók. Más szóval a hanghullámból vissza tudunk következtetni arra, hogy milyen volt a beszélő szervek állása a hang keltésekor. Ez pedig egy lehetséges első mozzanat az egyes hangok felismerése, az akusztikai lényegkiemelés felé.

Záró gondolat

A gépi beszédszintézis világszerte és hazánkban is rendelkezésre álló eljárásai sok régi szolgáltatás feljavítására és sok új szolgáltatás bevezetésére adnak módot a távközlésben. A magasabb rendű gépi beszédfunkciók — a beszédmegértés, beszélőazonosítás stb. — tekintetében ugyanez a közeljövőben várható.

IRODALOM

[1] *Gordos G., Takács Gy.*: Digitális beszédfeldolgozás, Műszaki Könyvkiadó, Budapest, 1983. p. 345.

[2] *Kempelen, W. v.:* Le Mechanisme de la Parole, suive de la Description d'une Machine Parlante, J. V. Degen, Vienna, 1791.
[3] *Dudley, H., Tarnóczy T.:* The Speaking Machine of Wolfgang von Kempelen, JASA, Vol. 22. 1950. pp. 151—160.
[4] *Fletcher, H.:* Speech and Hearing in Communication, van Nostrand, New York, 1953.
[5] *Reeves, A. H.:* Francia Szabadalom 52183, 1938.
[6] *Dudley, H :* Remaking Speech, JASA, Vol. 11. 1939. pp. 169—177.
[7] *O'Neill, E. F.: TASI:* Time Assignment Speech Interpolation, Bell. Lab. Rec., Vol. 37. March, 1959. pp. 82—87.
[8] *Gordos, G.:* Speech Detection in Severe Noise, 11st Int. Cong. on Acoustics, Paris, 1983. Proc. pp. 91—94.
[9] *Gordos, G.:* Digitalizálás a hangtechnikában: új távlatok az ember—gép kapcsolatban, Kép és hangtechnika, Vol. XXX. No. 1. Febr. 1984. pp. 15—23.
[10] DT 1000 Digitalker Speech Synthesis Evaluation Board, National Semiconductors IM—FL 30M120, 1980.
[11] MEA 8000, Philips gyártmányismertető, 1982.
[12] *Békési S., Gordos G., Olaszy G., Podoletz Gy., Takács Gy.:* Eljárás formánsszintetizátorok vezérlésére mesterséges beszéd és speciális hangjelenségek létrehozása céljából, Magyar találmányi bejelentés, 18 682, 1983.
[13] *Atal, B. S., Hanauer, S. L.:* Speech Analysis and Synthesis by Linear Prediction of the Speech Wave, JASA, Vol. 50. 1971. pp. 637—655.
[14] *Itakura, F., Saito, S.:* Speech Analysis-Synthesis System Based on the Partial Autocorrelation Coefficient, Acoust. Soc. of Japan Meeting, Oct. 1969.
[15] TMS 5200, Texas Instruments gyártmányismertető, 1982.
[16] *Gordos G., Podoletz Gy., Békési S., Takács Gy.:* Eljárás és berendezés a beszédkeltés akusztikus csőmodelljén alapuló beszéd és egyéb hangjelenségek mesterséges előállítására, Magyar találmányi bejelentés, 4186/1983.