Beszédinformációs rendszerek 2005 tavasz

Beszédkeltés gyakorlat (Olaszy Gábor, Kiss Géza)

A beszéd biológiai rendszerek között működik. A beszéd akusztikai jele egyedi és egyéni, pillanatnyi produktum eredménye. Kétszer nem tudjuk ugyanazt a mondatot ugyanolyan akusztikai szerkezettel kimondani, csak hasonlóval, amit a percepciós rendszer dolgoz fel olyan formában, hogy a mondanivalót megértsük. Ez a tény problémát okoz mind a beszédszintézisben, mind a gépi beszédfelismerésben. A szintézisben nehéz szép hang előállítása folyamatos és természetes dallammal, hangsúllyal (prozódiával). Kicsit mindig döcög a szintetizált beszéd. A gépi felismerésnél pedig a sokféle reprezentációból történő lényegkiemelés elvégzése jó hatásfokkal a gond.

A beszéd nyelvi szintjei: társadalmi (általában a magyar beszéd) és egyéni (a beszélő hangja)

A társadalmi szintet kell megvalósítani pl. egy beszédszintetizátornál (elvárás, hogy mindenki megértse)
Az egyéni szint feldolgozása a gépi beszédfelismerés egyik legnehezebben teljesíthető követelménye (mindenki beszélhet a géphez, ebből kell a lényeget, a kiejtett hangsort felismerni)

Alapfogalmak, meghatározások

1. A beszéd hangzó részek és szünetrészek sorozata. A hangzó rész hordozza az információt, a szünet rész engedi, hogy a hallgató felfogja az információt. A két rész egyformán fontos szerepet kap a beszédkommunikációban. Ha valamelyik hiányzik, vagy nem megfelelő szerkezetű, akkor sérül a kommunikációs hatékonyság. A szünetek hiánya például a gépi beszéd érthetőségét erősen rontja. A szünet detekciója a gépi beszédfelismerés egyik kulcskérdése.
2. A beszéd végeredménye (a hullámforma), sok paramétertől függ:
akusztikai, (milyen hangsor hangzik el)
nyelvi, (milyen információt hordoz a hangsor)
egyéni, beszélőtől függő (egyéni megformálás, hangszínezet, fiziológiai állapot, férfi-nő-gyerek beszél stb.)
3. A beszédkutatás és a beszédtechnológia művelése multidiszciplináris felkészültséget igényel! Nyelvész, informatikus, mérnök, fonetikus, orvos stb.

4. A beszéd elméleti szerkezeti szintjei

Szegmentális szerkezet (alap)	**szupraszegmentális szint (felépítmény)**
Az alapvető szerkezeti elemek a beszédben	**prozódia**
akaratunktól ált. nem függ	**akarattól függő**
részei: a beszédhangok és hangkapcsolódások	**részei: dallam (lásd a CD-n)**
akusztikai szerkezete	**hangsúly(lásd a CD-n)**
spec. Intenzitások (lásd a CD-n)	**ritmus (lásd a CD-n)**
spec. Időtartamok (lásd a CD-n)	**hangszínezet**

A hangok jelölése: V=magánhangzó, C= mássalhangzó

5. A beszéd akusztikai vizsgálatának építőelemei

frekvenciaszerkezet	**intenzitás-idő függvény**
Fo és a formánsok	**mondatra, szóra, hangra vonatkozó, ill. hangon belüli**

MINDEZEK AZ IDŐSZERKEZET KERETÉBEN VALÓSULNAK MEG.
Lásd részletesen az MNYBA CD-ről.

6. Alapfogalmak
Fo = alapfrekvencia= a hangszalagok nyitódásának, záródásának gyakorisága: 50-500 Hz. Ez adja a beszéd dallamát, része van a hangsúlyozásban is. Férfiaknál átlagosan 100 Hz, nőknél 200 Hz.

<u>Formáns</u>= az artikulációs csatorna rezonanciafrekvenciaáin felerősített felhangnyaláb. Jelölése: F1, F2, F3, F4, F5, Formáns sávszélesség: B1=30-50 Hz, B2=100-300 Hz, B3= 400-700 Hz
A formánsok hatóköre. F1= 250-800 Hz, F2= 500-2300 Hz, F3=200-4500 Hz.

Zönge= a gége szintjén létrejövő kvázi periodikus rezgés (forrásjel), amely a hangszalagoktól ered. A zönge frekvenciája az Fo, alapfrekvencia. Ez a vonalas spektrum alapharmonikusa.
Fojtott zönge= zöngés zár- és zárrés hangok zárszakaszi építőeleme. Formánsai nincsenek. b,d,g,gy,dz,dzs
Néma fázis= a zöngétlen zár- és zárrés hangok zárszakaszi építőeleme. Hang nem keletkezik a néma fázisban. p,t,k,ty,c,cs
Zárfelpattanás= zárhangok zárszakaszi eleme utáni hangrész, amely befejezi a zárhangot (lökéshullám).
Specifikus időtartam= a hangra jellemző alapidőtartam a hangkörnyezet függvényében
Lásd részletesen az MNYBA CD-ről
Specifikus intenzitás= a hangra jellemző alapvető intenzitásszint a többi hanghoz viszonyítva. Lásd részletesen az MNYBA CD-ről
Locus=a mássalhangzó artikulációját jellemző frekvencia koncentrációk összessége a frekvencia tengelyen (Főleg a zárhangokra vonatkoztatják a V F2-jéből következtethető ki).

7. FONÉMA= a sokféle beszédhangból általánosított elméleti egység, amely megkülönböztető funkciót tölt be a nyelv hangsorainak jelentéstartalmában. Például: bár – pár; már- vár.
A hosszú-rövid hangok is külön fonémaként kezelendők a magyarban. Például: hal - hall; sok – sók; tör – tőr. A szintézisben a hosszú-rövid magánhangzókat külön kell kezelni, a mássalhangzók esetén csak a rövid hang időtartamát kell megnövelni (hangfüggő a nyújtási eljárás).
8. Hanghelyzet. A beszédhangoknak különböző szerkezeti megvalósulásai lehetnek hangsorkezdő, hangsorbelseji és hangsorzáró helyzetben. Lásd részletesen az MNYBA CD-ről.

9. Artikulációs sebesség (hang/s). Lásd részletesen az MNYBA CD-ről
10. Beszédsebesség (hang/s) Lásd részletesen az MNYBA CD-ről
11. VOT= zönge kezdési idő a zöngétlen zárhang zárfelpattanásától a zönge megindulásáig mért idő. (Orvosi diagnosztikai eljárásokban is használható.) Értéke nyelvenként változik.
A magyarra: p=10-20 ms; t=20-30 ms; k=30-100 ms

Az emberi beszédkeltés

- A tüdőből kiáramló levegő 3 féle gerjesztést hozhat létre: (i) zönge, azaz alapvetően zöngés kváziperiodikus jel (zöngés hangokhoz), (ii) alapvetően turbulens áramlás, zaj jellegű (pl. rés hangokhoz) és (iii) a kettő keveréke (pl. kevert gerjesztésű rés- és zárrés hangokhoz).

A hangszalag rezgéseiből alakul ki a fojtott zönge is a **zöngés** zár- és zár-rés hangoknál a zárszakasz idején.
Speciális gerjesztési forma a csend, amelyet néma fázisnak neveznek (a **zöngétlen** zár- és zár-rés hangoknál). Gerjesztésnek tekintjük a zárhangok zárszakaszai után létrejövő lökéshullámot is.
A zönge kialakulása a gégében történik (fix hely), a turbulens áramlások kialakulási helye változó az artikulációs csatornán belül (gége, veláris terület, palatális, alveoláris, dentialveoláris, labiodentális pontok). Ezért tud létrejönni a kevert gerjesztésű hang.
Az ember hangképző szervei (tüdő, gégefő, garat-, száj-, orrüreg). A gége feletti rész neve toldalékcső, vokális traktus, artikulációs csatorna, rezonátorrendszer. Hossza átlagosan 17,5 cm.

- Bemutató ábra: a formánsok kialakulása **Lásd részletesen az MNYBA CD-ről.**

A formánsok kialakulásához kell a zönge, mint gerjesztő, forrásjel, valamint egy változó keresztmetszetű cső jelenléte, amelynek pillanatnyi átviteli függvénye határozza meg a rezonanciahelyeket a frekvenciatengelyen (ez a beszéd során folyamatosan változik). Az átviteli függvény maximum helyei képezik ki a formánsokat a zöngéből. A formánsok adják meg a hang formáját (á, o, ű stb.). A formánsok rfrekvenciaértékei függetlenek az alapfrekvenciától.

- Hangok szerkezete: frekvencia, intenzitás, időszerkezet **Lásd részletesen az MNYBA CD-ről.**
- Hangok tulajdonságai: táblázat a formánsfrekvenciákról, hasonlóságok (képzés helye, módja szerint), előrehatás/visszahatás **Lásd részletesen az MNYBA CD-ről.**

A mesterséges beszédkeltés és alkalmazási köre

Hol használunk gépi beszédet?
1. a vizuális visszajelzés mellé kiegészítő információ

2. az ember-gép kommunikáció természetesebbé tétele
3. információ nyújtása ott, ahol máshogyan nincs mód
4. az emberi fárasztó munka kiváltása (telefonközpontoknál, információs rendszereknél, stb.)

Az emberi és gépi beszédkeltés különbsége
Az emberi beszédkeltés párhuzamos folyamatok eredménye: több paraméteres függvény, amely egyazon időben valósul meg (hangképzés, hangerő, időzítés, hangsúlyozás, dallamformálás stb.).
A gépi beszédkeltésben kénytelenek vagyunk ezt a folyamatot több részre bontani: soros (esetleg párhuzamos) megvalósítás.

A gépi beszédkeltés módszerei

Kötött szótáras rendszerek, illetve szövegfelolvasók

A két megvalósítás megközelítési módjának összehasonlítása.

Kötött szótáras beszédszintetizátor

Adott, előre meghatározott üzenetek kimondására alkalmas. A kimondandó üzeneteket emberi felolvasásból készítik el.
Tervezési lépések:

- Az összes üzenet feltérképezése: állandó / változó üzenetrészek. A változókra kell koncentrálni, azokat kell akusztikailag hozzáilleszteni az állandó üzenetekhez, hogy együttesen is jól érthető információt kapjunk. „*A tegnapi leveleinek száma: 23.*"
- A bemondó által felolvasandó szövegek összeállítása. Fontos: a szintetizátor adatbázisában eltárolt mondatok és az adatbázis elkészítéséhez felolvasott mondatok **nem ugyanazok**. Meg kell vizsgálni: milyen vivőmondatba kell a tényleges (majdan elhangzó) üzenetet beágyazni? Cél: az elhangzó üzenet teljes hosszában (az üzenet állandó és változó részében is) jó legyen az intonáció (dallam, ritmus, intenzitás).
- Összefűzési szoftver elkészítése: egy szabályrendszert valósít meg (mit, mikor, mivel kell összefűzni).

Tervezési problémák:

- A rendszer később nehezen bővíthető: új felvétel szükséges, a bemondó hangja szükségképpen változik, ezért el fog ütni a korábban felvettekétől.

Kész rendszer üzembevétele esetén fellépő kérdések

- Külföldi rendszer átvétele, honosítása esetén: hogyan lehet magyarítani (hazai fejlesztésben meg lehet-e oldani)
- Műszaki megbízhatóság + akusztikai arculat (szépen beszéljen, a beszéd és szünet aránya jó legyen)
- Célszerű akusztikai / fonetikai szakértő alkalmazása az elembázis összeállítására

Számfelolvasó (példa tipikus kötött szótáras rendszerre)

A számelemenként való összefűzés korszerűtlen, rossz minőségű, szaggatott hangot eredményez.
Ma, ennek ellenére a legtöbb rendszerben ez működik (telefonszámok bemondása, árak, időpontok, dátumok felolvasása, banki információ szolgáltatása, stb.)
A korszerű megvalósítás alapelve: a számelemek fonetikai kapcsolódásának figyelembe vétele. Ehhez olyan hangfelvételt kell készíteni, amelyből minden számelem minden pozíciójára és kapcsolódásaira folyamatos akusztikai illeszkedés hozható létre.
A felvételhez össze kell állítani olyan számsorokat (400-500 többjegyű szám), amelyekben az összes számelem előfordul az összes lehetséges pozícióban (számcsoport eleje, közepe, vége) és az összes lehetséges hangzókörnyezetben, egynél többször, hogy több választási lehetőség legyen. Ebből kell kivágni a 200-250 db végleges számelemet, amiből a kívánt szám hullámformáját összefűzéssel megvalósítjuk.
Tagolás: A számokat a nyelvnek megfelelő szabályok szerint tagoljuk (312 = három száz tizen kettő; drei hundert zwölf). A vezérlő (összekapcsoló) algoritmus is nyelvfüggő.

Sajnos jelenleg Magyarországon ezt a korszerű technológiát csak a BME TMIT-en fejlesztett rendszerekben használják, az egyéb rendszerekre a korszerűtlen megoldás és az igénytelen tervezés a jellemző, melynek eredménye a döcögős, kiegyenlítetlen hang.
Ábrák, demonstráció: példa rossz és jó számfelolvasásra. Bemutató a Számok ’96 rendszerből.

Szöveg-beszéd átalakítók (TTS)

Bonyolult, sokrétű, több tudományterület ismeretét igénylő feladat. CSAPAT MUNKA!
Két komponensből épül fel: agy (szabályrendszer, tudásbázis, adatszintű előkészítés), majd a megvalósítás (hangelemek, adatbázis és jelfeldolozás)
Előzmények
1791 Kempelen Farkas 20 éves munkával megalkotja a világ első beszédkeltő szerkezetét. A gép ma az MTA nyelvtudományi Intézetében látható és működtethető (rekonstruált változat)
1916 Bánó Miklós mérnök beadja Budapesten a világ első szabadalmát "Tetszőleges szöveg reprodukálására alkalmas beszélőgép" címmel.
1939 Amerikában bemutatják az első elektronikus rezgőkörökkel működtetett angol beszélőgépet.
1982 Az MTA Nyelvtudományi Intézetében megalkotják az első magyar elektronikus beszédszintetizátort. A neve: HUNGAROVOX.

TTS SZINTÉZIS

A TTS rendszerek hangkeltő komponensének megvalósítási módjai:

- Formánsszintézis (a hangok formánsértékeit adják meg az artikulációnak megfelelő sűrűséggel)
- Hullámforma összefűzéses módszerek (emberi beszédből kivágott részleteket kapcsolnak össze)
- Korpusz alapú szintézis (nagy beszédkorpuszból válogatják ki az összefűzendő hullámforma elemeket)
- LPC szintézis (LPC paramétereket adnak meg 10-20-ms-onként és egy gerjesztőjelet)
- HMM alapú szintézis (nagy beszédkorpuszból tanítják be a HMM-eket, majd a szintézisnél a kialakított paramétervektorokból állítják elő a beszédet egy kódolóval)

FORMÁNSSZINTÉZIS
A formánszintézis jellemzői:
A beszédhangokra jellemző formánsértéket tároljuk. Az ezek közötti átmenetet mesterségesen hozzuk létre interpolációs hangszeletekkel. A folyamatosságot mikrohangszeletek (gyakorlatilag állandónak tekinthető rövid részek) egymás után helyezésével hozzuk létre.
Hagyományosan a megvalósítás számítógéppel vezérelt külső hardver használatával történt: gerjesztő jel (zönge + zaj opcionálisan) és szűrősor. Ma már szoftver szűrőket használnak a formánsok kialakítására. A jövő kutatási iránya az emberi beszédet egyre pontosabban leíró formáns szintézis.
Előnyei: kis adatbázisokból (1-2 kB) működtethető (főleg régebben volt jelentős), tág határok között változtatható hangjellemzők (rekedt, suttogó, mélyebb, magasabb, ének stb.).
Hátrányai: gépies hangzás;

Bemutató: Univoice, Multivox 4 formánsszintetizátorok

Hullámforma összefűzéses szintézis

Alapelve: megfelelő hullámforma-elemek összefűzése; esetleg jelfeldolgozással intonáció ráültetése
Elemek: diád (két félhangból álló hullámforma); triád (CVC elemre a C-ket elvágjuk a felénél, a V teljes egészében marad). Ez a legtöbbet alkalmazott technológia 2000 óta.
Előnyök: a hangzókat és a hangszínt a tárolt hangelemek tartalmazzák -> a természeteshez közel álló hangzás viszonylag könnyen megvalósítható. A triádos jobb hangminőséget eredményezhet, mint a diádos, gondos tervezésnél. Tipikus példa ilyen rendszerre a profivox szövegfelolvasó technológia (regények, hírekre stb.).
http://alpha.tmit.bme.hu/pub/PROFIVOX-TTS-demo/ hírfelolvasás, gyógyszerinformáció, regény.

Hátrányok:
Új hangszín létrehozásához teljesen új adatbázis készítése szükséges.
A jelfeldolgozás során minőségromlás jön létre.

Bizonyos hangzó kombinációkat esetleg csak közelíteni tudunk. Példák: kapj (ha nincs zöngétlen **j**); i**ng**, ze**ng.**
A diádos és triádos rendszereknél szükséges az elemek egymáshoz illesztésének kézi elvégzése (akusztikai csiszolás). Ez sok élő munkát igényel.
Bemutató:
Multivox 5: Westel Mailmondó (1999. november óta)
AT&T szintetizátor: a formánsszintézissel ellentétes irány: nem csak hogy a szegmentális szintű részeket nem számítjuk ki, hanem szupraszegmentális részeket is a természetes beszédből vesszük.
Tervezési lépések, problémák:

- A hangadatbázis (paraméter adatbázis) összeállítása. Döntés: hangdefiníció: mely hangokat realizáljuk; a tárolt elemek típusa: diád, triád (melyek), hosszabb elemek (NUS – Non-Uniform Synthesis)
  A hangadatbázis ellenőrzése, javítása: hangsebészet (példák bemutatása)
  Bemutató: hangsebészeti példák: sz, c, t; t - tt
- Adatbázis készítése
  Felolvasandó mondatok kiválasztása: olyan környezetbe kerüljenek a hangok, hogy minél kevésbé hasson rájuk a hangzó környezet. (A 'k' hang jó).
- Betű -> hang átalakítás megtervezése: szöveg előfeldolgozás (számok, rövidítések kifejtése); szöveg-hang átalakítás (kiejtési szótár, rendhagyó kiejtések); hang-hang átalakítások (hasonulások)
- Prozódiai rész megtervezése (dallamvonal, szünetek, hangsúlyok, időtartamok); szegmentális szintű (paraméter alapúnál), szószintű, mondat szintű, szöveg szintű intonáció
- Elem összeállítás, intonáció ráültetésének problémái (intonációs algoritmusok: LPC, PSOLA: az előadáson bővebben)

KORPUSZ ALAPÚ BESZÉDSZINTÉZIS
Jellemzők: nagy, több órányi beszédkorpusz kell ugyanazon beszélőtől, felcímkézve hanghatárokkal, szóhatárokkal, szünetekkel, hangsúlyokkal, hangszimbólumokkal.
Nagy szövegbázis kell, ami szoros szinkronban van a beszédadatbázis mondatainak címkéivel.
Költségfüggvényeket kell definiálni a válogatáshoz.
A beszédjel összerakásakor a szegmentális és szupraszegmentális szint nincs szétválasztva.
Előnyei: szép hang, de változtatni nem lehet. (a jövő kutatási iránya a benne rejlő lehetőségek miatt)
Hátrányai: Hibát javítani nehéz, nagy számítási kapacitást igényel, minden hanghoz nagy adatbázist kell készíteni, ami költséges és fárasztó, a hangkaraktert nem lehet változtatni. Csak kötött témakörhöz kapcsolódva lehet a szép hangot biztosítani (menetrend, időjárás, adott termékekre vonatkozó árlista, stb.).
Bemutató: http://alpha.tmit.bme.hu/pub/PROFIVOX-TTS-demo/ időjárásjelentés, menetrendi tájékoztató.

Minőségbiztosítás, tesztelés
Kész beszédszintetizáló rendszer választásakor mire figyeljünk:
Minőségbiztosításos mérési eredmények alapján való értékelés

- Jó beszédminőség (lásd a beszédszintetizátorok minősítése részben részletesen a CD-n)
- Elfogadható ár
- Terméktámogatás
- Alkalmazáshoz való illesztési lehetőség

Beszédszintetizátorok minősítése (lásd részletesen az MNYBA CD-n)

I. Szegmentális szintű minősítés. Az adott nyelv hangjait és a hangok egymásra hatását jól valósítja-e meg. Példa: vonószenekar, egészségtár, elmondtátok, színház, függvény, lesz,

II. Szupraszegmentális szintű vizsgálat. Az adott nyelv prozódiai követelményeit hogyan teljesíti. Mondatfajták dallama, hangsúlyozás egyszerű és összetett mondatokra. Különösen fontos az adott alkalmazásba ágyazva vizsgálni.

III. Az alkalmazáshoz kapcsolódó egyéb követelmények vizsgálata. Pl. Név és címfelolvasónál hogyan kezeli a nevek felolvasását.
Nevekre: Kováts=Kovács; Tymothy=Timóti; Pléh=Plé; Schumacher=Súmaher
Ericsson=Erikson; BMT Kft=Bé Emm Té Kft; 1+1 Bt= egy plussz egy Bt.

Címekre: Kossáthy u. =Kosáti utca; Czázár A. u.=Cázár András utca
A é. II.3/c= A épület, második emelt 3 per cé

Ajánlott olvasmányok angol nyelven:

- International Journal of Speech Technology 2000/3-4: beszédszintézissel kapcsolatos írások (Tanszéki könyvtár)
- Acta Linguistica Hungarica 2002 dec. (MTA könyvtár)
- G. Olaszy, G. Németh, G. Gordos: The MULTIVOX multilingual text-to-speech converter. In: Bailly, G.-Benoit, C.-Swallis, T. (eds): Talking Machines: Theories, Models and Applications. Elsevier-North-Holland Publishers. Amsterdam 1992, p385-411.

Ajánlott olvasmányok magyar nyelven:

- Beszédkutatás sorozatból. 2004-2008 MTA Nyelvtudományi Intézet, Kempelen Farkas Beszédkutató Lab.
- Olaszy G.: Elektronikus beszédelőállítás (1989)
- Olaszy G.: Gépi beszédkeltés információs rendszerekhez Magyarországon. Akusztikai Szemle, III. évf., 1-3 szám, 1999., 4-13. oldal
- Gordos-Takács: Digitális beszédfeldolgozás (Egyetemi könyvtár)
- Gósy Mária: Fonetika, a beszéd tudománya. Osiris Kiadó 2004.
- Olaszy Gábor: Hangidőtartamok és időszerkezeti elemek a magyar beszédben. Akadémiai Kiadó. 2006.
- Olaszy Gábor: Mássalhangzó-kapcsolódások a magyar beszédben. Tinta Könyvkiadó. 2007.

Linkek:

MULTIVOX4 magyar szövegfelolvasó formánsszintetizátor
ingyenes szoftver, letölthető: http://alpha.tmit.bme.hu/pub/multivox4
Beszédkutatási Labor, TMIT http://speechlab.tmit.bme.hu
Fonetikai vonatkozások, oktatási anyagok, Kempelen beszélőgépe
http://fonetika.nytud.hu
A magyar beszéd akusztikai szerkezetének bemutatása, spektrogramok, intenzitások, hullámformák, hangkapcsolódások stb.
http://fonetika.nytud.hu/cvvc