A jegyzet tartalmazza a VIK Wiki oldalán található BeszInfo vizsgák kérdéseit és válaszait 2009 tavaszáig bezárólag. Akad a végén ZH is, de sajna a legtöbb ZH-hoz nincsenek kidolgozva a válaszok, pedig van h a vizsgán ZH-ból vesznek kérdést. Ajánlom nézni már a 2002-es vizsga feladatait is, mert mindig előjön olyan feladat (pl 2010-ben) ami már volt nagyon régen, vagy ha nem is ugyan az, de akkor visszavezethető rá (pl a kreatív feladatok). By Jero ;)

2002.01.23. vizsga

1.a) Osztályozza a beszédhangokat a létrehozásukhoz használt gerjesztés szempontjából! Jellemezze a beszédhangokat (spektrális-, intenzitás-, idő-) szerkezetük szempontjából. (10 pont)

b) Mi az időablak szerepe a beszéd színképe elemzésében? Mi az előnye és mi a hátránya a rövid és a hosszú ablaknak? (8 pont)

Gerjesztés szerint lehet:

- zöngés: az összes magánhangzó, b, d, g, gy, v, j, m, n, ny, l, r
- zörejes: p, t, ty, k, c, cs, f, sz, s, j*, h
- kevert: dz, dzs, z, zs

Akusztikai szerkezet szerint:

- Egyszerű: az összes magánhangzó, v, f, z, sz, zs, s, j, h, m, n, l
- Összetett: b, p, d, t, g, k, gy, ty, c, cs, dz, dzs, ny, r

[Forrás: CD, 138. oldal]

Specifikus időtartamok:

- Magánhangzók: i,u,ü,o,a,e,ö,é, á (70 és 160ms között rendre)
- Mássalhangzók
  - 40ms: r
  - 50ms: n,l
  - 60ms: z, zs, réshangok
  - 70ms: p, t, k, ty
  - 80ms: f, sz, s
  - 90ms: c, cs

Intenzitás:

- Imin: h
- Imax: á,e
- Magánhangzók csökkenő sorrendben: á,e, a, é, ö,o, i, ü, u

[Forrás: CD, 93. oldal]

(Ha valaki tudja, pontosabban mire gondolhattak, az szerkessze át, gondoltam mégsem kérik az összes hang kifejtését egyenként...)

2. Pontokba szedve írja le a tervezés és megvalósítás menetét egy HI-FI számfelolvasó elkészítéséhez. (20 pont)

Kötöttszótáras rendszer ésszerű.

1. tematika meghatározása -> számfelolvasás
2. felhasználók osztályozása: felhasználó, laikus
3. üzenetek meghatározása: mivel HIFI minőség kell, így igényesebb rendszereknél ajánlott az elemet megelőző és követő hangelemeknek megfelelő változatait is letárolni pl gyezer meg nyezer, nem csak ezer (így is max pár100 elem lesz), állandó meg tartalom nincs a specifikáció szerint
4. felolvasandó szöveg megtervezése, vivőmondatok: utóbbiak nincsenek, tehát elég az összes lehetséges hangelemkapcsolódásnak megfelelő számosságú üzenetet kiválasztani
5. bemondó kiválasztása: mindegy, csak ne a palik
6. hangfelvétel, HI-FI minőségben!
7. digitalizálás, ügyelni kell arra hogy a hangminőség ne romoljon, megfelelő bitráta stb. megválasztása
8. adatbázis elkészítése, elemek kivágása
9. próbaüzem, akusztikai csiszolás
10. rendszerintegrálás

3. Egy 8 kHz-es mintavételi frekvenciával és az alábbi H(f) karakterisztikájú visszaállítóval működő mintavételező rendszer bemenetére a *sas*, majd a hangsor kerül egymás után férfi ejtésben állandó frekvenciával (F0: 125Hz).

$$H(f) = \begin{cases} 1, & \text{ha } 1 < |f| \le 3.5 \\ (4 - |f|) / 0.5, & \text{ha } 3.5 < |f| < 4 \\ 0, & \text{egyébként} \end{cases}$$

a) Megkülönböztethető-e a két visszaállított hangsor hangzása? Miért? (8 pont)
b) Mi változik, ha a rendszer bemenetére is egy H(f) karakterisztikájú szűrő kerül? (5 pont)
c) Javasoljon egy olyan mintavételi frekvenciát és összetett simító karakterisztikát, amely a fenti hangsorokat helyesen és elfogadható komplexitással megvalósítva átviszi! (7 pont)

F: zöngétlen réshang, nincsenek zörejgócok, egyenletes eloszlás a 1000-10000Hz frekvenciatartományban. A környezetében levő magánhangzó formánsaira csak kis mértékben van hatással.

S zöngétlen réshang: zörejelemek 1800-6500Hz között, intenzív zörejgóc ált. 2500-3500Hz között. Az s hangot követő magánhangzó formánsaiban kismértékű mozgás van jelen az átmeneti fázisban.

a) Metalogika alapján: a /c miatt nem a válasz! Indoklás: mivel mindkettő hang zöngétlen, azaz gerjesztése zörejes (fehérzaj szerű), ezért spektrumukban

mindenfele frekvenciakomponens előfordul, és egész magas frekvenciákon is vannak fontos komponensek, ezeket ez a mintavételezés (telefon) nem viszi át, ezért az "f" és "s" nehezen megkülönböztethető, a kis mintavételezési frekvencia miatt fellép az átlapolódás jelensége is.

b) Megszűnik az átlapolódás jelensége, az s zörejgóca így könnyebben kivehető és megkülönböztethető az f hang egyenletes frekvenciaeloszlásától. (ebben nem vagyok biztos)
c) 22khz mintavételezéssel, és egy darab hasonló szűrővel 1 és 11khz között (egyenletes meredekségű) a probléma megoldható.

Szerintem nem átlapolódásról van szó, hanem Aliasingról. A jeneség megszüntetése Anti Aliasing Filterrel lehetséges (egy aluláteresztő szűrő). ha a két fogalom ugyanazt jelentené, akkor én kérek elnézést... Aliasing def: Ez akkor lép fel, ha a mintavevő-tartóra ráengedünk a mintavételi frekvencia felénél nagyobb komponenseket is, amelyik így spektrális átlapolódásba kerülnek a hasznos jel periodikus spektrumával, és megjelennek olyan „ál" komponensek, amelyek az eredetiben nem voltak benne. -> ez szvsz ugyan az -- TitCar - 2007.05.25. Ez bizony ugyanaz -- Maco - 2010.01.06.

4. A vezetékes telefon frekvencia átviteli tartománya 300-3400Hz. Mely beszédhangok torzulnak el leginkább az átvitel során? Miért nem zavar ez minket a gyakorlatban? (10 pont)

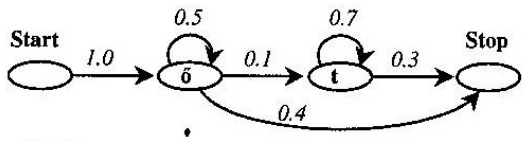
Leginkább a zár- és zárréshangok torzulnak az átvitel során, mivel a spektrális szerkezetükben ezek a hangok rendelkeznek nagyon magas frekvenciaösszetevőkkel (4khz fölött is), amit a telefonvonal szűrője levág. A gyakorlatban azért nem zavaró, mivel a hangok 4khz alatti komponensei is adnak némi támpontot, valamint az ember a magasabb értelmezési szinteken a hang- és szövegkörnyezetből is következtetni tud arra, hogy milyen hang lehetett ott.

(Más ötlet?)

5. Beszédfelismerési kísérleteket végeztünk lényegkiemelésként csupán a keretenkénti logaritmikus energiát számolva. Az alábbi HMM hálózatra 2 db egymás utáni jellemzővektor került. (Ezek nyilván nem származhattak valós szóbemondásból, inkább hibás szó-detekcióból.). Mit ír ki a felismerő, ha az állapotfüggő megfigyelési valószínűségsűrűség-függvények Gauss-fv.-ből állnak az alábbi paraméterekkel:

$m_{\text{„ó"}} = 5.0, \quad \sigma_{\text{„ó"}} = 1.0$

$m_{\text{„t"}} = 2.0, \quad \sigma_{\text{„t"}} = 1.0$



és a jellemzővektorok értéke: $o_1 = 4.0, \quad o_2 = 3.0$?

Válaszát részletezett számításokkal indokolja!

/Számolási segítség: Gauss-fv: $\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$; $\exp(-0.5) = 0.606$/ *(15 pont)*

Valószínűségek:

- Ő állapotban O1 felismerésére: 0.242
- Ő állapotban O2 felismerésére: 0.054
- T állapotban O1 felismerésére: 0.054
- T állapotban O2 felismerésére: 0.242

2 lehetséges út (felismerés) van, valószínűségük:

- START-Ő-T-STOP: 1.0 * 0.242 * 0.1 * 0.242 * 0.3 = 0.00176, azaz 0.17 %
- START-Ő-Ő-STOP: 1.0 * 0.242 * 0.5 * 0.054 * 0.4 = 0.00261, azaz 0.26 %

vagyis a felismerő *"őő"*-t ír ki.

6. Témavezető informatikusként a következő feladattal bízzák meg: Web, WAP és telefonos lekérdezési felületet kell megvalósítani a budapesti és a londoni részvénytőzsde értékeihez. Milyen főbb beszédtechnológiai elemeket kellene ill. lehetne alkalmazni a rendszerben? Milyen tervezési lehetőségeket kellene figyelembe venni? Gondolkozzon kreatívan és széles látókörűen! A kérdésekre több jó válaszegyüttes is adható! (17 pont)

WAP:
mivel az adatátvitel szűkös beszédátvitelhez (adatforgalom+beszéd?), ezért lehetne az, h sms-ben elküldik a kért tőzsdeinformációkat, aztán az sms-t felolvassa az SMSmondó
vagy eleve egy telefonon futó kliensalkalmazás leszedi a szükséges adatokat a szerverről WAPon keresztül és olvasná fel így integrált módon egymaga oldaná meg a problémát
mindkét esetben egy egyszerű beszédszintetizátor jöhet csak szóba a mobiltelefonok (mondjuk azt hogy) egyelőre még szűkös kapacitása miatt, tehát formáns vagy diádalapú megoldás az elsődleges jelölt.
mindkét esetben a kliensprogram nyelvét a felhasználó telepítésnél választhatná ki

Telefon:
Itt már a vezérlés is lehet, sőt ajánlottan beszédalapú. Gondosan megtervezett, dialógusszerű, 2-3 hierarchiaszintes szerkezetben közölhetné igényét a felhasználó, pl: Mire kíváncsi? Értékpapír. Melyik cég papírja? Machester United. Milyen mutatójára? Értékére. 4.52 penny.
vagy kötetlen kérdésfeltevés után kérdezne rá a program a bizonytalan vagy hiányos részletekre: Hogy állnak a Machester papírjai? 4.52 penny.
a válaszok vegyes (TTS+kötött) felolvasó rendszer segítségével generálódnának, dinamikus rész lenne a számok, dátumok, esetleg cégnevek felolvasása, a többi statikus, vivőmondatok
ügyelni kell hogy többnyelvű legyen a rendszer, hisz ez mégis a világ legnagyobb tőzsdéje, alapból lehetne angol, és ha nem reagál a felhasználó, akkor mondaná adott nyelveken hogy melyik gombot nyomja meg ha héberül/kínaiul/xu! vagy egyéb nyelven szeretné hallani a frankót
a beszédfelismerő beszélőfüggetlen legyen, a beszédgenerátor triádosként lenne optimális, esetleg diádos

a hang minősége 3,7khz, ezt mind beszédgenerálásnál (minták minősége), mind beszédfelismerésnél figyelembe kell venni

PC:
számítógéppel a billentyűzet, de főleg az egér legalább olyan gyors kommunikációt biztosít mint a beszéd, kivéve kereséseknél
tehát adott, kevés alternatíva közül egér segítségével, sok (cégek neve, dátum keresése) lehetőségnél beszéddel választanánk
a felismerő futhatna a felhasználó gépén, csökkentve a szerver terheltségét, ugyanakkor nem szabad túl nagynak lennie, valamint az sem jó ha be kell tanítani minden szóra (DTW kiesett), tehát HMM jön szóba, moderált méretű paraméteradatbázissal, esetleg beszélőadaptív lehetne!
a beszédgenerátort mindeképp a felhasználó gépén lenne érdemes megoldani, ahol alapértelmezetten/adott parancsra felolvasná a lekérdezés eredményét
a program nyelvét a felhasználó választaná ki telepítésnél, de a kliens természetesen nyelvfüggetlen módon, csak adatok formájában kommunikálna a szerverrel

(további ötletek nyugodtan jöhetnek, ez egy kreatív feladat!!)

2002.05.23 vizsga

1. Magyarázza meg a következő fogalmakat.
a) Formáns. Adjon példákat
b) Gerjesztés és fajtái. Adjon példákat.
c) Egyszerű és összetett szerkezetű beszédhang. Adjon példákat.
d) Erősen, gyengén ls kölcsönösen illeszkedő beszédhangok. Adjon példákat.
e) Prozódia. Milyen komponensei és alkomponensei vannak? (3 – 3 pont)

a) **Formáns**

- Röviden: A zöngejelbol az artikulációs csatorna üregrendszere által felerosített felhangnyaláb. [forrás: Beszéd CD, 7.oldal]
- Hosszabban: A zöngés beszédhangok létrehozásához két független építőelemre van szükség: a gerjesztő jelre (zönge, alaphang, alapfrekvencia: Fo) és az artikulációs csatornára, amelyik a zönge jelét átformálja. Az átformálás során a zönge adott felharmonikusait az üregrendszer rezonanciái felerősítik. Ezeket a felerősített felhangnyalábokat formánsoknak nevezik. [forrás: Beszéd CD, 84.oldal]
- Például az "a" hang formánsai: F1 500-600Hz, F2 900-1100Hz, F3: 2200-2400Hz

b) **Gerjesztés**: A beszédhangok létrehozásának egyik alaptényezője a gerjesztés, vagyis a hangforrás, amiből az artikuláció hatására a tényleges beszédhang kialakul. A gerjesztési hang alapvetően háromféle lehet: zöngés, zörejes és kevert. [forrás: Beszéd CD, 62.oldal]

- zöngés: az összes magánhangzó, b, d, g, gy, v, j, m, n, ny, l, r
- zörejes: p, t, ty, k, c, cs, f, sz, s, j*, h
- kevert: dz, dzs, z, zs

c) **Egyszerű** szerkezetű a beszédhang, ha időben periodikus vagy állandó.
Összetett szerkezetű, ha belső időszerkezettel is rendelkezik, ez írja le a hangon belüli akusztikai jelenségek időtartam-értékekeit, azok egymáshoz viszonyított időarányait.

- Egyszerű: az összes magánhangzó, v, f, z, sz, zs, s, j, h, m, n, l
- Összetett: b, p, d, t, g, k, gy, ty, c, cs, dz, dzs, ny, r

d) Artikulációjuk a beszédhangok szerint lehetnek **stabilak** (nem illeszkednek a környező magánhangzókhoz), **erősen illeszkedők** (nagyban befolyásolja akusztikai jellemzőket, formánsaikat a szomszédos beszédhang) és **kölcsönösen illeszkedők** (a szomszédos hanggal kölcsönösen befolyásolják egymás jellemzőit, paramétereik közelednek).

- Stabil: gy, ty, j, n, ny, r
- Kölcsönösen illeszkedő: b,p, d, t, dz, c, dzs, cs, v, f, z, sz, zs, s, h, m, l, az összes magánhangzó (?)
- Erősen illeszkedő: g, k

e) **Prozódia:** A prozódia a beszéddallam, a hangsúly, a ritmus, a hangero, a tempó és a hangszínezet nyelvi használata, a beszédképzés szupraszegmentális szintjének része. [forrás: Beszéd CD, 7.oldal]
Komponensei:

Dallam
Hangsúly (ezen belül alkomponensek: alapfrekvencia, intenzitás, időtartam)
Ritmus
Hangszín (?)

2.

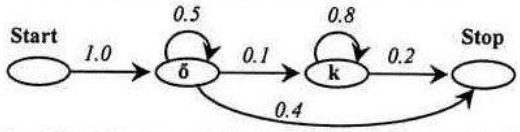
a) Milyen szempontok szerint lehet minősíteni a beszédfelismerő rendszereket? (10 pont)

- statisztikai alapú (HMM, ANN) vagy szabálybázis/tudásalapú
- beszélőfüggetlen, beszélőfüggő vagy adaptív (avagy beszélők száma alapján)
- akusztikus környezet alapján: robusztus (zajos környezetben is használható), távbeszélő minőséggel vagy kiváló hangminőséggel működik csak
- szociolingvisztika: dialektusra, korra és nemre érzékeny e
- artikuláció alapján: izolált szavas, kapcsolt szavas vagy folytonos (diktáló) rendszer
- szótárméret: kis (<100 szó), közepes vagy nagy (>20.000 szó)
- beszédstílus: spontán, parancsmódú vagy dialógus-menüszerű
- nyelvfüggés-nyelvazonosítás
- alkalmazói környezet: szakembereknek vagy laikusoknak, egyfelhasználós vagy sokfelhasználós

b) Miért rejtett a rejtett Markov modell? (5 pont)

A modell azért rejtett, mert egy megfigyelés esetén nem lehet egyértelmuen meghatározni, hogy melyik állapot generálta azt. [forrás: Beszéd CD, 7. oldal]

3. Beszédfelismerési kísérleteket végeztünk, lényegkiemelésként csupán a keretenkénti logaritmikus energiát számolva, Az alábbi HMM hálózatra 2 db egymás utáni jellemzővektor került.



(Ezek nyilván nem származhattak valós szóbemondásból, inkább hibás szó-detekcióból.). Mit ír ki a felismerő, ha az állapotfüggő megfigyelési valószínűség-sűrűség függvények Gauss-fv.-ből állnak az alábbi paraméterekkel:

$m_{\text{„ő"}} = 5.0, \quad \sigma_{\text{„ő"}} = 1.0$

$m_{\text{„k"}} = 2.0, \quad \sigma_{\text{„k"}} = 1.0$

és a beérkezett jellemzővektorok értéke: $o_1 = 4.0, \quad o_2 = 3.0$? Válaszát részletezett számításokkal indokolja!

/Számolási segítség: Gauss-fv, $G(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$; $\exp(-0.5) = 0.606$, $\exp(-1.0) = 0.367$, $\exp(-2.0) = 0.135$/

(20 pont)

Kis bevezető:
A HMM feladatok lényege, hogy a START állapotból a STOP állapotba vezető utak közül meghatározzuk a legvalószínűbbet. Egy út valószínűsége nem más, mint az út során meglépett átmenetek valószínűségének és az az út által bejárt állapotokban megfigyelt események valószínűségének a szorzata. Tehát ha elindulok a START állapotból, és először az "ő", utána pedig a "k" állapotba érek, a következő módon történik az út valószínűségének a számítása:
átmenet STARTból "ő"be: 1.0

- "ő" állapotban az O1 eseményt megfigyelni: 0.242
- "ő" állapotból "k" állapotba lépni: 0.1
- "k" állapotban O2 eseményt megfigyelni: 0.054
- "k" állapotból a STOP állapotba érni: 0.2

Ezek szorzata pedig 0.00267, azaz az út valószínűsége 0,267%

Triviális dolgok de inkább leírom mégegyszer:

- Az út felváltva áll lépésekből és megfigyelésekből
- A START állapotból a STOP állapotba kell eljutni
- MINDEN eseményt meg kell figyelni valamely állapotban, azaz annyi köztes állapotot kell érinteni START és STOP között, ahány megfigyelt vektorunk vam
- Ezen belül az összes lehetséges út közül kell kiválasztani a legvalószínűbbet

Általános esetben dinamikus programozás szerű módszerrel lehet a legjobb utat megtalálni (táblázat kitöltése, alulról felfelé a megfigyelt események szerint sorban haladva vízszintesen, felfele meg a lehetséges állapotok. Balra nem léphetünk, mert újra nem figyelünk meg eseményt, a simán eggyel jobbra való lépés meg a helybenmaradással ér fel, stb, ha az utolsó eseménynél nem értük el a STOP állapotot, az nem út, a STOP állapotig elért utak közül a legjobbat választjuk ki), Viterbi algoritmus segítségével, de ezek annyira egyszerű példák h ránézésre megmondod azt a 2 utat ami lehetséges, kiszámolod melyik mennyi, és az nyer aki előbb..

Megoldás:

Valószínűségek:

- ő állapotban O1-t megfigyelni: 0.242
- ő állapotban O2-t megfigyelni: 0.054
- k állapotban O1-t megfigyelni: 0.054
- k állapotban O2-t megfigyelni: 0.242

Két út lehetséges, ezek valószínűsége:

- START-Ő-K-STOP: 1.0 * 0.242 * 0.1 * 0.242 * 0.2 = 0.00117, azaz 0.117 %
- START-Ő-Ő-STOP: 1.0 * 0.242 * 0.5 * 0.054 * 0.4 = 0.00261, azaz 0.267 %

Tehát "őő"-t ír ki a felismerő.

Ezt még annyival egészíteném ki - szerintem egyáltalán nem triviális, bár ki lehet sakkozni -, hogy a megfigyelést úgy kapod, hogy a vizsgasoron megadott Gauss-függvénybe behelyettesítgetsz, m lesz az állapotok/betűk m-je, x pedig az o megfigyelésvektorok. Tehát ha annak a vg-t akarod kiszámolni, hogy az állapotban -et figyeled meg, akkor a függvénybe -t és -t helyettesíted be.

-- Zsolti - 2007.05.30.

4. Számítsa ki a suttogott beszéd (átlagos hangnyomás 1000 mikroP) és a kiabálás (átlagos hangnyomás 1000000 mikroP) közötti dinamikatartományt dB-ben. Mi a jellegzetes különbség a suttogó és a normál beszéd spektrális színképe között? (15 pont)

Suttogó beszéd: 1000 * 10ˇ-6 P = 10ˇ-3 P

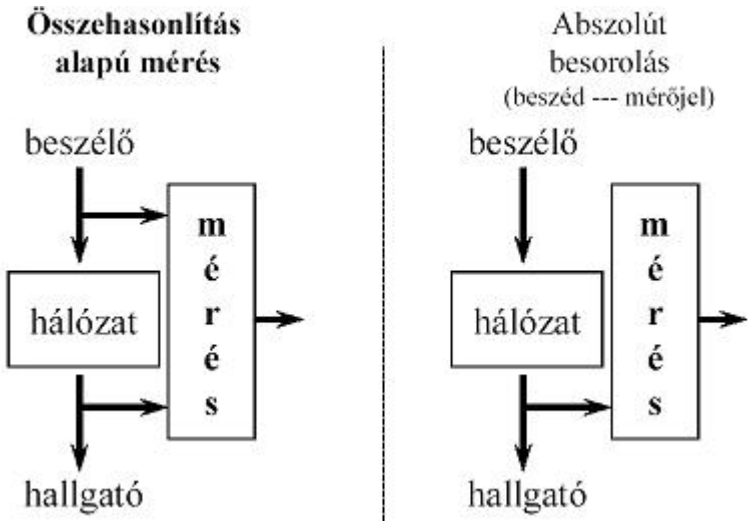
Kiabálás: 1000000 * 10ˇ-6 P = 1P

Dinamikatartományuk különbsége: 20 * log (1P / 10ˇ-3 P) = 60dB (mivel 1000x hangosabb, ez is a jó eredmény -- 20*3 az 60)

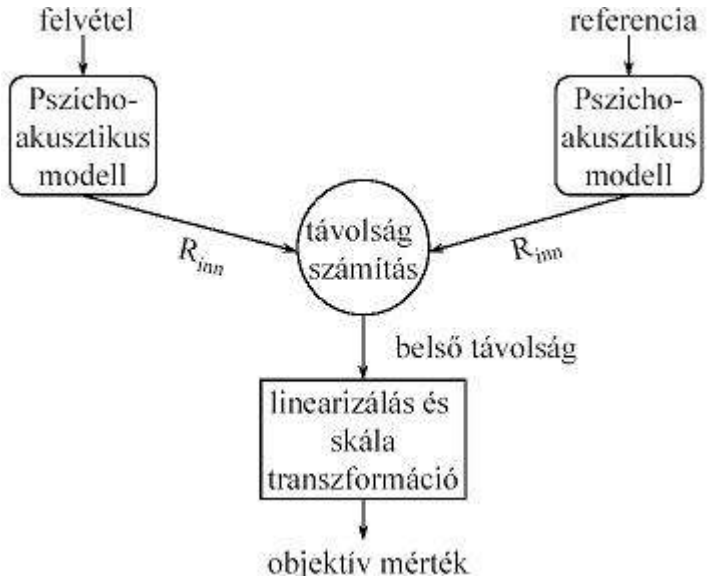
Különbség abban rejlik hogy suttogó beszédben nincs zöngés gerjesztés, így alapfrekvencia és formánsok sem, tehát a magánhangzók vonalas színképe helyett is folytonos színképet kapunk spektrális elemzésnél. (ehhez még sztem lehetne írni..)

5. Ismertesse egy gépi beszédminősítő működési elvét:
a) adja meg a blokkvázlatot, (8 pont)
b) röviden ismertesse minden elem szerepét (7 pont)

Beszédminősítő rendszer blokkvázlata I.: összehasonlítás alapú mérés:



Beszédminősítő rendszer blokkvázlata II.: a jelfeldolgozás általános menete:



A hálózat az az átviteli technológia vagy kódolás, amelynek beszédminőségét minősíteni szeretnénk. A mérés menete a második ábrán van kifejtve, ezen belül a pszichoakusztikus modell az, ami az ember számára lényeges, hallható részek kiemelése és a nem érzékelhetőek, észlelhetőek elnyomása, majd ezekből ablakozással és Fourier transzformációval jellemző vektrokat ad ki keretenként. A távolságszámítás ezek alapján a vektorok alapján megbecsüli a forrásjel és a vett jel eltérését, torzulását, amelyből linearizálás és skálatranszformáció után kapjuk meg a vizsgált rendszer beszédminőségének objektív mértékét.

(Ez így korrekt?)

6. Tervezze meg egy <u>ésszerű</u> BKV menetjegy-bérlet vásárló rendszer párbeszéd folyamatát, ami egyaránt működik beszédfelismerővel és nyomógombos vezérléssel is. A beszédfelismerő <u>egyszerre</u> maximum 10 különböző szót képes felismerni, kb. 95%-os biztonsággal. (Elsődleges input a beszédfelismerő!) A rendszert 6-99 éves korig bárki használhatja (külföldieknek is legyen esélye!). A készüléknek egysoros kijelzője van, billentyűzet 10 gombos. (20 pont)

Vásárolható:

- **Jegy (normál, gyűjtő, ..)**
- **Bérlet (30 napos, havi, éves …)**
- **Turista jegy**
- **…**

- első lépéskent a 10 gombbal indítja el a vásárlást az utas, minden gomb egy-egy nyelvnek felel meg, így levesszük a nyelvazonosítás terhét a rendszerről és tudja egyből, mikortól van action (esetleg azzal indíthatnák el, hogy bemondják a kívánt nyelvet: Magyar, English, Deutsch, stb)
- minden, amit a gép mond, sorban kiíródik a kijelzőn is!!
- a nyelv kiválasztása után üdvözöli a rendszer az utast napszaknak megfelelően majd megkérdezi mit szeretne venni, vagy információt kérni
- ha nincs response, akkor elkezdi mondani a gép hogy miket lehet venni: (vonal)jegy, szakaszjegy, átszállójegy, gyűjtőjegy, turistajegy, (havi) bérlet, 30 napos bérlet, éves bérlet, stb.
- a választ a következő formában várja a rendszer: "[Kérek/Szeretnék venni/Adjon/Aggyá] [1-10] (Vonal)jegyet/átszállójegyet/szakaszjegyet/gyűjtőjegyet/turistajegyet/([diák/nyug díjas](havi)/30 napos/éves bérletet) [kérek/szeretnék (venni)]", ezek nagyjából max. 10 hosszú mondatoknak felelnek meg, lehetőleg a kulcsszavakra vadászva, azokra forszírozva illesztünk
- ha pontatlan a kérés (diák v nyugdíjas bérlet?) akkor rákérdez a rendszer és megvárja a releváns választ
- mindezen fix választási opciók esetén (tehát nem dátumnál vagy darabszámnál) a 10 gombbal is lehet választani az alternatívák közül
- a felhasználó minden ponton a "vissza" paranccsal egy kérdéssel visszaugorhat (ha rosszul értette meg a rendszer), "elölről" paranccsal pedig újrakezdheti a vásárlási procedúrát
- ha a végleges kérést megértette a gép (úgy véli) akkor visszamondja, mit szűrt le és felszólítaná az utast h dobjon be ennyi meg ennyi pénzt vagy vissza vagy elölről ha nem tetszik valami. Pl.: "Két darab vonaljegyhez dobjon be 300 forintot. Ha meggondolta magát vagy félreértettem, mondja vagy nyomja hogy vissza vagy kezdjük elölről" (A biztos hatás érdekében ez a két parancs gombbal is legyen kiadható) A gép szövegelését bármikor meg lehet szakítani pénzbedobással vagy gombnyomással.

- a pénz bedobását hanggal is nyugtázná a gép (bár ez lehet zavaró, tehát ha 5 mpig nem dob be új pénzt, és nem elég akkor szólalna meg hogy hello, ez így sovány)
- ha elég pénzt bedobott az utas, akkor megköszöni a gép, kiadja a jegyeket és a visszajárót, megköszönné hogy igénybevették a szolgáltatását, stb.

2006.05.22. vizsga

1. Mondjon 3 – 3 példát arra, hogy milyen tényezők okozzák az akusztikai paraméterek variáltságát, egy személyen belül és a személyek között. (12 pont)
Személyen belül:
- érzelmi állapot (nem sikerült beszédvizsga, lediplomáztam, lemerült a telóm)
- egészségügyi állapot (rekedt, megfázott, csuklik)
- szituáció (családi ebéd, szónoklat, történetmesélés)
- ...

Személyek közötti:
- nem (női hang magasabb, ffi mélyebb)
- ritmus (hadar, dadog, megfontolt)
- beszédhibák (selypít, raccsol)
- ...

2.
a) A hangkapcsolódások osztályozásánál az egymással kapcsolódó hangokra milyen kategóriákat lehet megkülönböztetni? Adjon példákat. (5 pont)

- Erősen illeszkedő: pl a g,k hang erősen illeszkedik a szomszédos magánhangzóhoz
- Kölcsönösen illeszkedő: sok hang, pl b,p,d,t kölcsönösen hatással van a szomszédos magánhangzókkal egymás formánsmozgásaira
- Stabil: gy, ty hangra kevés hatással van a környezet

b) Melyik magyar beszédhangok tartalmazzák a legmagasabb frekvenciakomponenseket és ezeknek mi a jellemző frekvencia tartománya? Miért fontos tudni távközlési alkalmazásokban? (5 pont)

- dz, c: 4000-5000 5500-7000Hz
- dzs, cs: 3700-8000Hz
- v, f : 1000-10000Hz
- z, sz: 4000-4500, 5000-8000Hz
- zs, s: 3700-8000Hz

a mintavételezés és a helyes antialiasing-szűrő megválasztásánál szükéges tudni a zörejgócokat és frekvenciatartományokat, hogy érthető maradjon a beszéd az átvitel során

c) Az y=F2, x=F1 koordináta rendszerben helyezze el az *i, u, á* hangokat. Rajzolja fel a rájuk jellemző spektrumot, ha feltételezzük, hogy F0= 200 Hz. (5 pont)

- iF1: 250-350Hz, iF2: 2300-2500Hz
- uF1: 250-350Hz, uF2: 500- 600Hz
- áF1: 700-800Hz, áF2: 1300-1400Hz

(ebből remélem mindenki fel tudja rajzolni egy két dimenziós térben a hangokat + a spektrumot)

d) Milyen lenne a spektruma ezeknek a hangoknak, ha suttogva ejti ki a beszélő. (5 pont)

a formánsszerkezet megmaradna, de a suttogás miatt a teljes frekvenciatartományban megjelennek kisebb komponensek, a spektrumképen a teljes frekvenciatartomány kicsit "beszürkülne".

Különbség abban rejlik hogy suttogó beszédben nincs zöngés gerjesztés, így alapfrekvencia és formánsok sem, tehát a magánhangzók vonalas színképe helyett is folytonos színképet kapunk spektrális elemzésnél.

(ehhez még lehetne sztem írni)

3. Igaz/Hamis (lusta vagyok)

4. Adja meg a következő rövidítések jelentését és válaszoljon a feltett kérdésekre.

a) Mi az a DTMF? Van-e szerepe a beszéd értehetőségében? Azonosítható-e és hogyan a jel spektrumában (5 pont)
Dual Tone MultiFrequency, DTMF jelek esetén nincs beszédjel, így zavarja az érthetőséget, mert 2 szinusz hang szólal csak meg, így a jel spektrumában könnyen felismerhető lesz a 2 kiugró amplitudó

b) Mi az F2 ill. B2? Hogyan határozhatóak meg? (5 pont)
F2 a beszédjel második formánsa, avagy az akusztikum második legkisebb felerősített felhangnyalába, a B2 pedig ennek a formánsnak a sávszélessége. F2 meghatározható a jel spektrumából, ez a második legkisebb erősítési hely (lokális maximum), a B2-t pedig ezen a maximum alatt 3 dB-lel meghúzott vonal és a burkológörbe metszéspontja jelöli ki.

c) Mi az ITU p.800? A beszéd mely jellemzőire vonatkozik? (5 pont)
ITU P.800: az ETSI egyik szubjektív beszédminősítő szabványa. Minősíthetünk

- abszolút módon, előre definiált skála alapján (ACR)
- 'jelenség' észlelési tesztek
- romlás megfigyelése eredetihez képest (DCR)
- referencia rendszerrel összehasonlítás (MNRU)

d) Mi a VXML, a SAPI és SUI kapcsolata? (5 pont)
Mindegyik a beszédinformációs rendszerek felépítését segíti, illetve annak egy eleme.

- VXML: Voice eXtensible Markup Language (http://en.wikipedia.org/wiki/VXML) - dialógusok tervezését segítő leírónyelv
- SUI: Speech User Interface, avagy beszédalapú felhasználói felület
- SAPI: Speech Application Programming Interface (http://en.wikipedia.org/wiki/Speech_Application_Programming_Interface) - a Microsoft beszédalapú felhasználói felület API-ja. Ezzel még nem dolgoztam, de például Symbian-ban van egy tts() függvény, amibe csak berakod a stringet, és a telefon elvégzi a beszédszintézist

5. Egy mai teherautóba épített beszédfelismerővel működő navigációs rendszert többen is szeretnénk használni. Milyen specifikációs feltételekkel lehetséges ez?

- Nem kiváló a hangminőség, robosztus rendszer kell
- Nem lehet emiatt diktáló rendszer, maximum kapcsolt szavas felismerő
- Beszélőfüggetlen kell legyen
- Előzőek miatt kis-közepes szótárnagyság a reális
- A rossz körülmények miatt fel kell készíteni spontán beszéd felismerésére
- Egyértelműen statisztikai alapú felismerő jön szóba (ilyenek működnek is, rossz a hangminőség és sok a beszélő)
- Mivel a GPS-nek ez nem a fő funkciója, fontos szempont hogy olcsó legyen a megvalósítása
- Ne kelljen a túlzott társszükséglet miatt növelni a készülék fizikai méreteit (bár nem tudom hogy ez ma még felmerülhet-e egyáltalán)

6. a) Tervezze meg egy telefonos, magyar nyelvű, magyarországi egyetemi felvételi információs rendszer párbeszédfolyamatát és sorolja fel specifikált beszédtechnológiai elemeit. A rendszer egyaránt működik beszédfelismerővel és nyomógombos vezérléssel is. Az izolált szavas beszédfelismerő egyszerre maximum 500 különböző szót képes felismerni, kb. 90 %-os biztonsággal. A rendszert legalább 10 éves időtartamra lényegi módosítás nélküli megoldással tervezze meg.

Szükséges információk:

- **Milyen szak(párok)ra kíváncsi**

Kimenet:

- **Az adott szó(párok) indító egyeteme(i) és kara(i) neve, címe, felvételi feltételek, pontszámítás és korábbi évek adatai**

Gondolkozzon kreatívan és széles látókörűen! A kérdésekre több jó válaszegyüttes is adható! (13 pont)

- Egy gombbal lehet indítani a rendszert, ezzel együtt esetleg nyelvet is ki lehet választani, így nem kell nyelvadaptációt és beszéddetekciót végeznünk
- A felhasználót megfelelően üdvözöli a rendszer, majd megkérdezi hogy milyen szakra, szakpárra kíváncsi

- A következő bemenetet várja: [A] {szak}/{szakpár} [ra/re vagyok kíváncsi]/[után érdeklődök] (a kérdéssel jól behatároltuk az adható válasz formáját!)
- Amennyiben a rendszer nem biztos a szakban, felsorolná a 10 legvalószínűbb szakot, amit mondhatott a felhasználó, és felszólítaná h válasszon közülül vagy mondja be újra
- A felismerő HMM alapú, robosztus, közepes nagyságú, beszélőfüggetlen.
- A válasz következő formátumban generálódna: A {szak}/{szakpár}t a következő egyetemek indítják: ([egyetem], [kar])*
- Ha a felhasználót nem érdekli az adott egyetem, "tovább" vagy "vissza" szavakkal léptethet (gyorsabban) közöttük (gombbal is)
- Ha felkelti érdeklődését valamelyik, a "címe", "felvételi (feltételek)", "pontszámítás", "korábbi évek" paranccsal kérheti le az őt érdeklő adatokat a karról (gombbal is választhat)
- Cím esetén: "A(z) {egyetem} {kar} címe [város] {közterület neve} [közterület] {házszám}, {irányítószám}
- Felvételi feltételek: [adott kar feltételei], a paraméterszerű adatokat dinamikusan generálja
- Pontszámítás: [pontszámítás menete], paraméterszerű számokat, adatokat dinamikusan
- Korábbi évek: [évben] [a ponthatár] {szám} [pont volt, a felvettek száma] {szám} [fő, a jelentkezők száma:] {szám} [fő] stb.
- Vegyes felolvasó rendszert használunk: TTS+kötött
- a [] elemek kötöttszótáras módon, előre felvéve vannak letárolva, a {} részek generálása pedig triádos szövegfelolvasó rendszer feladat*a
- A felhasználó a "lista" paranccsal tér vissza a megfelelő egyetemek listáájához (gombbal is)
- "Köszönöm" esetén vagy 1 perc tétlenség után a rendszer alaphelyzetbe áll

b) Milyen főbb szempontokat kell figyelembe venni, ha a feladat a spanyol vagy a szlovén felvételire vonatkozna az adott ország nyelvén? (4 pont)

- Szórendre, dátumra, számok felolvasására kell figyelni
- Teljesen más lehet a felsőoktatás menete, pontszámítási módszerek, ezeket is megfelelően át kell alakítani
- A felismerés során más sorrendben adja meg az adatokat a felhasználó
- Más temperamentumú beszéd, más beszédstílus: újra kell paraméterezni a felismerőt, nem csak a felismerendő szavakat kell kicserélni
- A spanyolok szieszẗáznak (koradu. lekapcsolhat a rendszer) //ez csak poén, senki ne vegye komolyan!
- ...?
- A katalán és a spanyol nyelv eltér, érdemes a nyelvek közé mindkettőt felvenni (spanyol rendszer esetén)

2006.06.02. vizsga A Csoport

1. Minden alkérdésre egy-két szavas válaszokat kérünk ebben a feladatban! (3 – 3 pont)
a) Milyen jellel mérjük a beszédátviteli rendszerek minőségét?
Természetes emberi beszéddel, de érdektelen felvételeket kell felolvastatni az alanyokkal! (nem vagyok teljesen biztos h ezt kérdezik..)
b) Az objektív minősítő rendszer hatékonyságát mihez képest mérjük?
Az objektív minősítés célja a szubjektív minősítés közelítése, tehát azt nézzük, hogy mennyire egyezik az eredménye az egyéni véleményekkel.
c) Ha a gépi minősítés a szubjektív minősítéshez képest egyes méréseknél lényegesen jobb, más méréseknél lényegesen rosszabb eredményt ad, akkor a minősítő mely komponensét kell módosítani?
A pszichoakusztikus modellt, esetleg a belső távolság számításának a módszerét (amivel a referenciafelvételtől való eltérést mérjük, számítjuk)
A csomagkapcsolt beszédátviteli rendszerek (pl. VoIP) mely tulajdonsága okozza a legnagyobb nehézséget a beszédminőség mérése során?
(A hálózat paramétereinek nem stabil volta. Teljesen más minőséget kapunk ha kis illetve szélessávon mérünk, illetve változatos kapcsolat (műholdas, kábel, adsl) esetén is jelentős eltéréseket tapasztalhatunk a beszéd minőségében, a hálózatforgalmi szituációkat nem is említve (pl. ha közben töltünk is).) Nem a rizsára voltak kíáncsiak. A válasz: Jitter (késleltetés-ingadozás). Bővebben: jegyzet

2.
a) Mikor és ki készítette az első beszédkeltő gépet a világon? Hol látható?
Kempelen Farkas, *1791*-ben. Ma az MTA Nyelvtudományi Intézetében látható. (legalábbis ajánlom neki h ottlegyen..) !! Update: "Az egyetlen megmaradt példány ma a müncheni Deutsches Museumban van." Forrás: http://hu.wikipedia.org/wiki/Kempelen_Farkas
b) Mikor és ki adta be a világ első szabadalmát tetszőleges szöveg felolvasására alkalmas beszélőgépre?
Bánó Miklós, *1916*-ban.
c) Mi az artikulációs sebesség? Milyen érték jellemző a magyarra? Mi a beszédsebesség?

- Az artikulációs sebesség az időegység alatt ejtett hasznos beszédhangok száma folyamatos ejtésnél, szünetek nélkül.
- A magyar beszédnél tipikus értéke 13 hang/s.
- A beszédsebesség a beszéd hangzásának teljes idejében, szünetekkel, időegység alatt elhangzott beszédhangok száma, a nem hasznos beszédjeleket is beleértve. (Magyar beszédnél 14 hang/s)
- artikulációs sebesség <= beszédsebesség

d) Mi a VOT? A beszédjel mely részén mérhető? Adjon 5 konkrét példát indoklással!

- VOT: Voice Onset Time avagy zöngekezdési idő
- felpattanó zárhangok esetén a zár felpattanása és az azt követő magánhangzó megszólalása között eltelt idő

- Tipikusan a beszéd azon helyen mérhető, ahol gerjesztésváltás történik, és zöngétlen hangot zöngés hang követ.
- A fentiek fényében a VOT pl. p után 8ms, t után 15ms, k után 26ms. (Ide lényegesen többet nem tudok írni, főleg az indoklás részét nem értem)

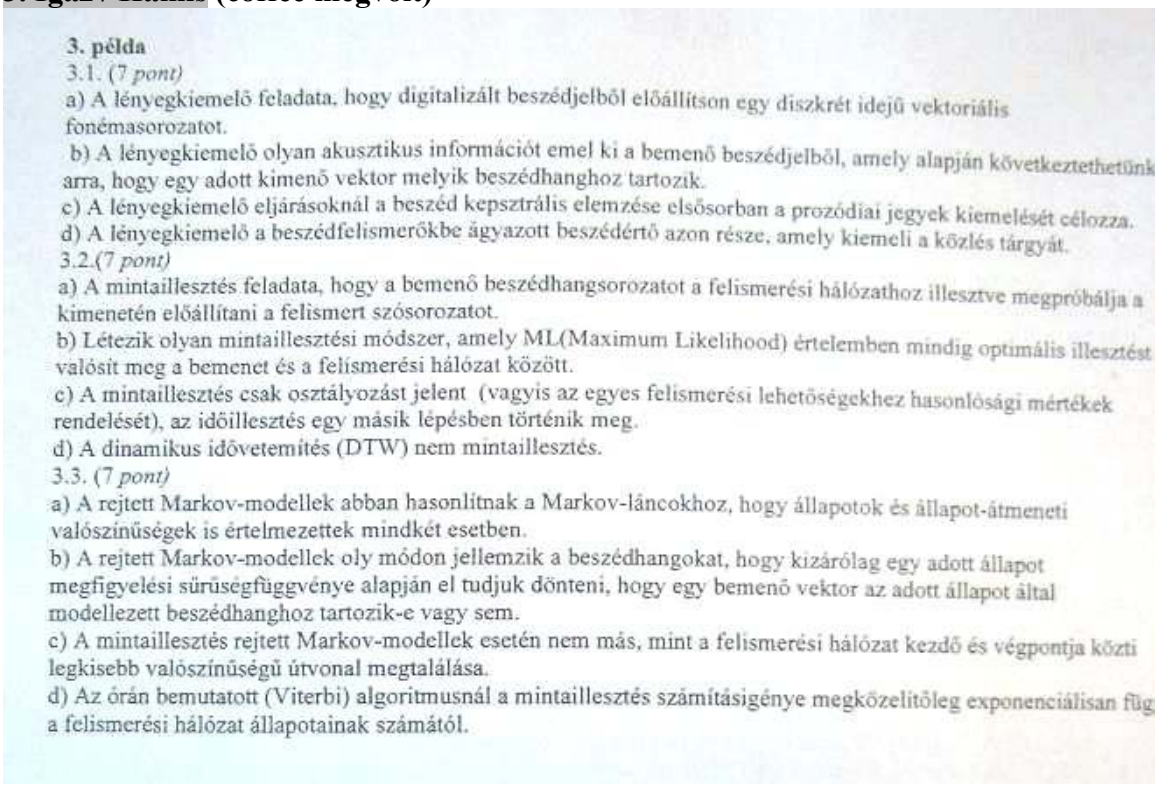
e) Mi a spektrális átlapolódás oka mintavételezéskor? Hogyan előzhető meg? Adjon példát.

- Spektrális átlapolódás: ha a hang mintavételezésénél a mintavételezési frekvencia kisebb, mint a legnagyobb frekvenciakomponens kétszerese, a visszaállításnál nemkívánatos jelek kerülnek visszaállításra, a jel nem állítható elő egyértelműen/hűségesen.
- Megelőzhető megfelelő karakterisztikájú aluláteresztő szűrővel a bemeneten. (Sávkorlátozás)
- Példát mindenki remélem tud adni ezek alapján :]

f) Mi a néma fázis? Sorolja fel az összes beszédelemet, amelyre vonatkozhat!

- Néma fázis: A zárhangok azon része, amelyben nincs hangképzés. A tüdőből kiáramló levegő a toldalékcsőben képzett akadály miatt feltorlódik és a zárfelpattanásig levegőáram nem hagyja el az artikulációs csatornát.
- A fentiek alapján néma fázis található a zöngétlen zár- és zárréshangoknál így: p, t, k, ty, c, cs.

3. Igaz / Hamis (coffee megvolt)



3. példa
3.1. (7 *pont)*
a) A lényegkiemelő feladata, hogy digitalizált beszédjelből előállítson egy diszkrét idejű vektoriális fonémasorozatot.
b) A lényegkiemelő olyan akusztikus információt emel ki a bemenő beszédjelből, amely alapján következtethetünk arra, hogy egy adott kimenő vektor melyik beszédhanghoz tartozik.
c) A lényegkiemelő eljárásoknál a beszéd kepsztrális elemzése elsősorban a prozódiai jegyek kiemelését célozza.
d) A lényegkiemelő a beszédfelismerőkbe ágyazott beszédértő azon része, amely kiemeli a közlés tárgyát.
3.2.(7 *pont)*
a) A mintaillesztés feladata, hogy a bemenő beszédhangsorozatot a felismerési hálózathoz illesztve megpróbálja a kimenetén előállítani a felismert szósorozatot.
b) Létezik olyan mintaillesztési módszer, amely ML(Maximum Likelihood) értelemben mindig optimális illesztést valósít meg a bemenet és a felismerési hálózat között.
c) A mintaillesztés csak osztályozást jelent (vagyis az egyes felismerési lehetőségekhez hasonlósági mértékek rendelését), az időillesztés egy másik lépésben történik meg.
d) A dinamikus időtvetemítés (DTW) nem mintaillesztés.
3.3. (7 *pont)*
a) A rejtett Markov-modellek abban hasonlítnak a Markov-láncokhoz, hogy állapotok és állapot-átmeneti valószínűségek is értelmezettek mindkét esetben.
b) A rejtett Markov-modellek oly módon jellemzik a beszédhangokat, hogy kizárólag egy adott állapot megfigyelési sűrűségfüggvénye alapján el tudjuk dönteni, hogy egy bemenő vektor az adott állapot által modellezett beszédhanghoz tartozik-e vagy sem.
c) A mintaillesztés rejtett Markov-modellek esetén nem más, mint a felismerési hálózat kezdő és végpontja közti legkisebb valószínűségű útvonal megtalálása.
d) Az órán bemutatott (Viterbi) algoritmusnál a mintaillesztés számításigénye megközelítőleg exponenciálisan függ a felismerési hálózat állapotainak számától.

3.1

a) HAMIS. Nem fonémasorozatot kell előállítania, hanem egy olyan 10-40 dimenziós vektort, melyeknek kicsi az intraindividuális és az interindividuális jellemzője.

b) IGAZ.
c) HAMIS. Semmi köze a prozódiához, a beszéd kisebb egységeinek kezelésében segíti munkánkat.
d) HAMIS. No comment

3.2
a) IGAZ. Kicsit furán van megfogalmazva, de szerintem jó.
b) IGAZ.
c) HAMIS. A mintaillesztés egyik lényege hogy a különböző ritmusú ejtések között is tudjon mintailleszteni.
d) HAMIS. Igaz csak sablonalapú és a legegyszerűbb fajta, de mintaillesztési eljárás.

3.3
a) IGAZ.
b) HAMIS. Valószínűségekkel dolgozik a HMM, így teljes biztonsággal sosem tudja megmondani, hogy egy megfigyelés adott állapothoz tartozik vagy éppen nem tartozik.
c) HAMIS. Legnagyobb valószínűségi útvonalat keres.
d) HAMIS. Mert lineárisan, lásd dinamikus programozás.

4.
a) Mi az LPC? Van-e szerepe a beszédértésben? Kapcsolatba hozható-e és hogyan a jel spektrumával?

- Linear Prediction Coding / Coefficients. Lineáris elõrejelzés. Olyan matematikai eljárás, amellyel a megelõzõ mintákból jósolni lehet a következõ mintát. LPC segítségével az akusztikus jelbõl meghatározható például az artikulációs üregrendszer átviteli karakterisztikája is.
- ??? (Ha jól meghatározhatók az LPC együtthatók, jobban érthetők a hangok?) A formánsokat jól lehet vele követni.
- Igen, a LPC analízis is egyfajta spektrumát adja meg a jelnek. (ide még lehetne írni)

b) Mi az F0 ill. F1? Hogyan határozhatók meg?

- F0 az alapfrekvencia, azaz a hangforrás gerjesztésének frekvenciája. F1 pedig a legkisebb (első) formáns azaz felerősített felhangnyaláb.
- F0 meghatározható a zöngés hangok periódusidejéből (megegyezik azokkal). F1 pedig a jel spektrumára illesztett burkológörbe első (lokális) maximumhelye.

c) Mi a Hamming-ablak és mi a szerepe a beszédfeldolgozásban?

- A Hamming-ablakot a jelre illesztve egy véges időtartományban kell csak elvégezni a Fourier-integrálást. A szerepe az, hogy adott időpillanatban releváns frekvenciákat felerősítse, a távoliakat gyengítse hogy adott időpillanatra jó spektrumot kapjunk a Fourier-integrálás után.

d) Mi a screen reader és a TTS kapcsolata?

- A screen reader csak egy illesztő alkalmazás a képernyő és a TTS között, a képernyőn található információt adja át felolvasásra a TTS számára.

5. Adjon meg min. 5 specifikációs szempontot egy távközlési szolgáltató számára tervezett e-levél felolvasó rendszerhez! Adjon meg min. 5 felhasználási lehetőséget is!

5 specifikációs szempont:

- Nyelv
- Operációs rendszer
- Beszéd minősége : érthetőség, természetesség
- Milyen hangokon szólaljon meg (ffi/női)
- Mennyire legyen paraméterezhető: hangmagasság, sebesség, szünetek hossza, stb.
- Vezérlési felület, API
- Bővítési, továbbfejleszthetőségi lehetőségek
- ...

5 felhasználási lehetőség:

- Emailek felolvasása telefonon keresztül
- Vakok és gyengénlátók számára
- Rendszerüzenetek, ajánlatok természetesebb közlése
- Előfizetési információk közlése emailen keresztül
- Gyerekek számára
- Call Center IVR (telefonos menürendszer) elemeinek dinamikus létrehozása, esetleg nagy kiterjedésű hiba esetén az 'üdvözlőszöveg' amiben bemondják hogy tudnak a hibáról és javítás alatt van, felolvasó nélkül beállítható
- ...

6. Sorolja fel a gépi beszédfelismerők jellegzetes fajtáit működési elv szerinti, használati módja szerinti, és méret szerinti osztályozásban.

Működési elv:

- Szabálybázisú
- Statisztikai alapú: HMM, ANN
- Sablon alapú: DTW (Dynamic Time Warping)

Használat módja:

- Spontán beszéd (folyamatos beszéd, pl diktáló rendszerek)
- Parancsmódú vezérlés (izolált szavas)
- Dialógusvezérlés (kapcsolt szavas, a szavak közötti szünetek minimálisak)

Méret:

- Kicsi: párszáz szó
- Közepes
- Nagy: 20-80 ezer szó

2006.06.02. vizsga B Csop (csak az eltérő kérdésekre kitérve)

3. Adjon meg min 5 specifikációs szempontot egy távközlési szolgáltató számára tervezett SMS felolvasó rendszerhez! Adjon meg min. 5 felhasználási lehetőséget is!

A szempontok kb ugyanazok, a felhasználási lehetőségek:

- Előfizetési információk természetesebb közlése
- Vakok és gyengénlátók segítése
- Idős felhasználók segítése, akik nem tudnak/akarnak kis képernyőn olvasni
- Autóval való közlekedés során is elolvashatjuk SMS-einket
- Email-eket SMSben továbbítva, azokat elolvashatjuk
- Minden olyan helyzetben előnyt jelenthet, amikor a nyomógombok használata vagy a kijelzőn megjelenő szöveg olvasása nem megoldható.

4.

a) Mi a SAMPA? Van-e szerepe a beszédértésben? Kapcsolatba hozható-e a jel spektrumával?

- SAMPA: Speech Assessment Methods Phonetic Alphabet. Beszédhangok jelölése 7 bites ASCII karakterekkel.
- A SAMPA-val a beszédhangok egyértelműen leírhatók, segíthet a beszédértésben.
- Szerintem nem hozható kapcsolatba a jel spektrumával. Vagy csak nagyon összetett, indirekt módon.

c) Mi a négyszögletes ablak és mi a szerepe a beszédfeldolgozásban?

A Fourier-integrálás során egy kis időkeret analízise úgy történhet meg, hogy az időben folyamatos jelet egymással átlapolódó négyszögletes ablakokkal kiablakozzuk. Így kis időszakaszokra megkaphatjuk a jel spektrumát, ami a magasabbrendű beszédfeldolgozás fontos alapeleme.

d) Mi a triád? Előnyei? Hátrányai? Mennyi egy nyelv lefedéséhez szükséges elemszám?

- Triád: Olyan hangkapcsolat, amelyben a középső hang egészben, a két szélső pedig részben van jelen. Beszédszintézisnél használják, elsősorban a magánhangzók szerepelnek középső helyzetben.

Előnyei:

- A magánhangzóknál nem lép fel torzítás a formánsok megtörése miatt.
- Természetesebb hangzás
- Könnyebb szövegtervezés

Hátrányai:

- Sok munkát jelent a felvétel
- Sok memóriát foglal
- Sok szöveget kell felolvastatni
- Diádokat és egyéb elemeket is igényel az adatbázis

Szükséges elemszám: , ennél némileg kevesebb mivel nem fordul elő minden hármas + a szükséges diádok: (szerintem a tisztán triádos adatbázis egyszerűen a fonémák köbével arányos. Az már a kevert adatbázis ahol diádok is vannak. Vagy?)

2007.05.25. vizsga

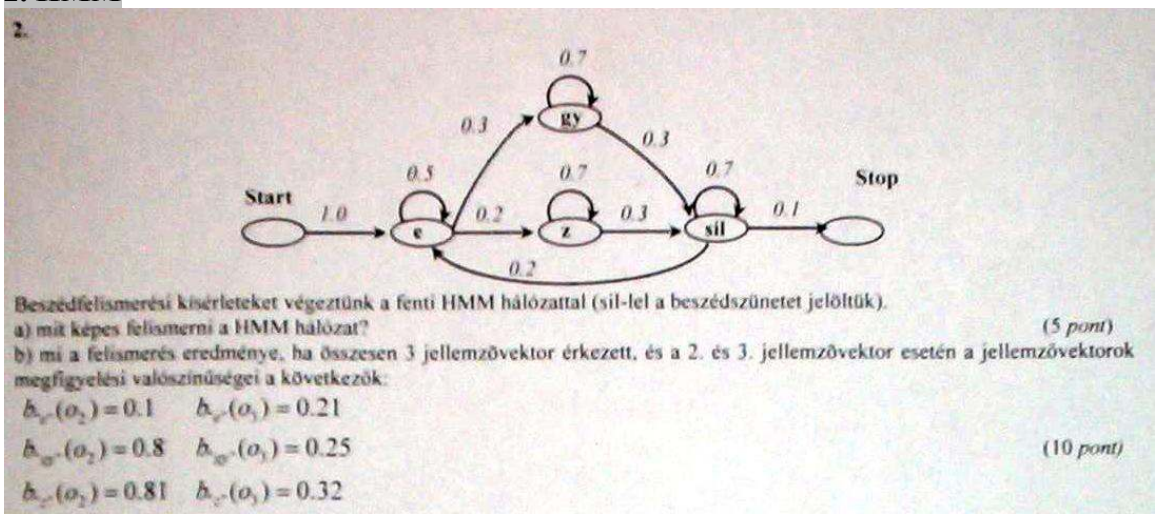
1. 11,025 kHz mintavételi frekvenciával, 16 bites lineárisan kvantált digitalizált beszéd felvételeink vannak. Spektrális elemzésre 256 pontos FFT-t számolunk (egy spektrum kiszámításához ennyi mintát használunk fel).

a) Mekkora lesz a spektrális elemzés legjobb idő-felbontása és jel/zaj viszony értéke?
Idő-felbontás: 256 pontos, és 11,025 KHz --> 90,7 mikroSec , innen az időfelbontás: 256 * 90,7 mikroSec = 23,2 msec.

SNR=1,74+n*6,02=1,74+16*6,02=98,06 dB

b) Mely beszédhang-csoportok spektrális vizsgálatát tudjuk és melyikét nem tudjuk ezekkel a felvételekkel lényegileg pontosan elvégezni?
Azokat nem tudjuk, melyeknek lényeges frekvenciakomponenseik vannak 5,5 kHz fölött, így például a zár és zárréshangok jó részét nem tudjuk így spektrálisan vizsgálni. Azért nem, mivel a mintavételezési frekvencia túl kicsi. Mint tudjuk, a mintavételezési frekvenciának 2x nagyobbnak kell lennie a legnagyobb frekvenciaösszetevőnél, így 11kHz esetén az 11/2=5,5kHz a legmagasabb frekvencia, amiket még jól tudunk mintavételezni, az ennél magasabbak átlapolódnak.

2. HMM



A táblázat nem túl jól olvasható, szerintem az első sorban 'e' utána 'gy' végül 'z' van az alsó indexben.

a) A HMM regexp szerűen felírva az ([egy|ez]sil)+ hangsort képes felismerni, azaz tetszőleges számú, de legalább egy "egy" és "ez" szót tetszőleges sorrendben, köztük szünetekkel.
b) Feltehetjük hogy az 1. jellemzővektor mibenléte érdektelen a számunkra, mivel azt mindenképp az "e" állapotban figyeltük meg, és ennek a valószínűsége közös minden más felismerése esetében, így csak a 2. és 3. jellemzővektor ill. az ezutáni bejárt

utak/állapotok döntik el, mi a legvalószínűbb útvonal. Szóval az "e" állapotban vagyunk és most következik 2 jellemzővektor. Mivel minden lépés után egy állapot és egy megfigyelés következik, valamint a 3. jellemzővektor után a STOP állapotba kell jutnunk, 2 további útovnal jöhet szóba: "z[sil]" illetve "gy[sil]". Ezek valószínűsége:

- z[sil]: Átlépés a "z" állapotba: 0.2. "z" állapotban O2 megfigyelése: 0.81. Átlépés a [sil] állapotba: 0.3. [sil] állapotban vektor megfigyelése: X. [sil] állapotból STOP állapotba lépés: 0.1. Összesen: 0.2 * 0.81 * 0.3 * X * 0.1 = 0.00486 * X
- gy[sil]: Átlépés a "gy" állapotba: 0.3. "gy" állapotban megfigyelése: 0.8. Átlépés a [sil] állapotba: 0.3. [sil] állapotban vektor megfigyelése: X. [sil] állapotból STOP állapotba lépés: 0.1. Összesen: 0.3 * 0.8 * 0.3 * X * 0.1 = 0.0072 * X

Összegezve: 0.0072*X > 0.00486*X, tehát a megfigyelés eredménye "egy[sil]".

3. 800 Mbyte kapacitású CD lemezen (44,1 kHz mintavételi frekvencia, sztereó felvétel, 16 bites lineáris kvantálás) állnak rendelkezésre egyenként átlagosan 3 perc hosszú zeneszámok. Szeretnénk belőlük csengőhangot készíteni egy olyan mobiltelefonra, ami 11,025 kHz-es mintavételi frekvenciával tud mono, 8 bites, A-törvényű logaritmikus kvantálású mintákat lejátszani és 16Mbyte szabad memóriája van.

a) Ábrákkal illusztrálja az átalakítás folyamatát! (8 pont)

Ábrák helyett az egyes lépések (kis dobozkákat rajzolnék egymás után, bennük az egyes lépések neveit írnám):

- Visszaállítom a kvantált, mintavételezett jeleket (sztereó!) analóggá.
- Átlagolom a két jelet időtartományban, amplitudó szerint 1 mono jellé.
- Aluláteresztő szűrő, mely 5 kHz-ig engedi át a jelet, persze 5 kHz körül lineáris gyengítéssel.
- Mintavételezés 11,025 kHzen.
- Kvantálás 8 biten.

b) Hány zeneszám van a lemezen? Valamennyi zeneszám átalakítható-e? Ha nem, mi lehet a megoldás? (6 pont)

1 sec hanganyag tárigénye: 44,1kHz mintavételezés, 16 bit, sztereó hangsávok: 44100*16*2= 1,411,200 bit = 172kbyte. 800Mbyte/172kbyte= 4763, azaz 4763 sec hanganyag tárolható, ami kb 79 perc. Ez 3 perces zeneszámokkal számolva 26 zeneszám. Nem alakíthatók át azok a számok, mely 5kHznél magasabb frekvenciakomponenseket tartalmaznak. Megoldás erre a fentebb már említett aluláteresztő szűrő.

c) Vissza lehet-e állítani az eredeti felvételt a telefonos formából? Ha igen, hogyan? Ha nem, miért nem? (6 pont)

Nyilván nem lehet visszaállítani a telefonos formából, ennek több oka is van. Egyrészt a monó hang átlagolással készült a sztereó hangsávokból, ezt lehetetlen visszaszűrni. (2 és 6 átlaga 4. 4 melyik két szám átlaga?). Másrészt az alacsony mintavételezés miatt elvesztjük az 5kHz feletti komponenseket, ezeket sem tudjuk visszanyerni. Harmadrészt pedig a 8 bites logaritmikus kódolás nem arányos a lineáris 16 bitessel, ezért főleg a magasabb tartományokban nagyobb lesz a kvantálásból eredő zaj nagysága.

4.

a) Mi a teljesítmény sűrűség spektrum, az akusztikai dB és a Phon érték kapcsolata?

- Az akusztikai dB-ből visszakövetkeztethetünk a hangjel amplitudójára (10-es hatványraemelés), az így kapott időjel négyzete a teljesítmény sűrűség spektrum. (ha jól mondom :])
- A Phon görbe pedig az azonos hangosságérzetű görbék serege, ahol a referencia-frekvencia az 1 kHz. Azaz 1kHz-es hangok esetén a Phon érték megegyezik az akusztikai dB-el.

b) Mi a Hanning-ablak és a szonogram kapcsolata?

- Ha gördülő spektrumot avagy szonogramot szeretnénk készíteni, akkor az időben folytonos jelünket bizonyos kis szeletekben mintavételeznünk kell. A kis kivágott időintervallumokból akkor kapunk jó spektrumot, ha azt megfelelően kiablakozzuk és nem csak simán kivágjuk egy négyzetes ablakkal. Egy ilyen jól bevált ablakozó függvény a Hanning ablak, melynek képlete:
  $0.5 – 0.5 * \cos(2\pi * t/T)$

c) Mi a VXML, a SUI és a DTMF kapcsolata a beszédinformációs rendszerekkel?

- Mindegyik a beszédinformációs rendszerek felépítését segíti, illetve annak egy eleme.
- A VXML avagy Voice eXtensible Markup Language interaktív dialógusok leírását és tervezését könnyíti meg ember és számítógép között.
- A SUI avagy Speech User Interface az ember-gép kapcsolatot beszéd és hangok által teremti meg.
- A DTMF avagy Dual Tone Multi Frequency egy jeltovábbítási megoldás avagy mechanizmus a normál telefonvonalon keresztül, ahol 2 frekvencia együttes megszólaltatásával összesen 16 különböző jelet generálhatunk (4*4=16).

d) Mi a locus, az F2 és F0 kapcsolata?

- A CV átmenet jellegzetessége a locus: megfigyelték, hogy pl. a d után ejtett magánhangzók felfutó szakaszait, ha visszafelé meghosszabbítjuk, ezek egy pontban metszik egymást – a legtöbb mássalhangzó az őt követő magánhangzó vagy őt megelőző magánhangzó második formánsát (F2) a szóban forgó mássalhangzót jellemző frekvenciára kényszeríti, ezek a locusok.
- Az F2 pedig nem más, mint a hangszalagoknál képzett gerjesztő jel alapfrekvenciájából (F0) a vokális traktusban felerősített, második legkisebb felhang-nyaláb (Fn).

5. Egy kötött szótáras telefonos információs rendszert kell terveznie egy áruház üzleti nyitva tartásának automatikus bemondására hetes időszakra. Csütörtökön az üzlet 20 óráig van nyitva, egyébként 18 óráig. Szombaton 11 óráig. Specifikálja a beszédtechnológiai alrendszereket és tervezze meg az információs rendszer dialógusát. Állítsa össze a felolvasó alrendszerben az építőelemek tárát úgy, hogy a koartikulációs hatásokat is figyelembe veszi a hullámforma összefűzésnél. Sorolja fel, hogy milyen elemeket fog tartalmazni az elemtár. Rajzolja fel az információs rendszer blokkvázlatát.

Először meg kell tervezni, hogy mit kell pontosan felolvasni a rendszernek. A leírás annyira kötött hogy a legegyszerűbb lenne egy egyszeri felvétel, mely szépen egy hanganyagban tartalmazná az összes információt. Ez nyilván elég merev lenne, másrészt

nem tennénk eleget abbéli kívánalmakban, miszerint kötött szótáras, telefonos rendszert kell készítenünk. Ekkor érdemes úgy megtervezni a rendszert, hogy információt fogadni is tudjon avagy egy beszédfelismerő modul is szükségeltetik mindehhez. Az információkérés avagy dialógus nagyjából így tervezhető meg:

- Üdvözlő szöveg, a végén kérdéssel, hogy melyik nap nyitvatartására kíváncsi a telefonáló. Ez egy fix szöveg.
- Ügyfél válasza, melyben a hét napjait (hétfő..), relatív utalásokat (ma, holnap) illetve konkrét dátumot (május 29) keresünk.
- A válasz értelmezése után esetleg visszakérdezés, ha nem értettünk semmit, esetleg DTMF-es megoldáshoz való folyamodás
- Válasz generálása egy mondatba ágyazva, a következő opciókkal: Az üzlet (ma/holnap ... hétfőn/kedden ... január 29-én) (szám) órától (szám) óráig tart nyitva.

A beszédfelismerő lehetne egy HMM-s rendszer pár szóra (kis szótár) minél robusztusabban (zajra érzéketlen, beszélőfüggetlen) betanítva. A következő szavakat kéne felismernia: hétfő-vasárnap, ma-holnap-holnapután-tegnap-tegnapelőtt, hónapok, 1-én ... 31-én. Ezt most nem is részletezem mert sztem nem erre kíváncsiak.

Beszédszintetizátor tervezése: A fix vivőmondat adott, a változtatandó részek: időpontok (ma/holnap, hétfőn-vasárnap, január-december, 1-én-31én) illetve számok (0-24-ig). Az időpontokat elég egyszer felvenni hiszen a mondatban csak egy helyen szerepelnek, viszont a hónap-nap kapcsolatokban előfordulhatnak bizonyos kivételek, amelyekre figyelni kell, bár most nem találtam ilyet (vki?). A számokat viszont kétszer kéne felvenni, mivel két pozícióban is szerepelnek (hangsúly, prozódia!), viszont nincs belőlük olyan sok (25 szám) ezért nem kell vacakolni a még kisebb egységekre bontással.

Innentől meg a szokásos szövegek elkészítése - bemondó kiválasztása - felvétel - tárolás - csiszolás - rendszerintegrálás blabla, meg valami ábra a fenti elemeket összefűző ábrával. Ne felejtsük itt el az értelmezőt és a szabályok alapján való elemkiválasztást!

6.

a) Mi a lényeges különbség a felhasználás szempontjából a beszélő-függő és a beszélő független beszédfelismerők között?

A beszélőfüggetlen rendszereket bárki, bármikor használhatja előzetes betanítás nélkül, viszont általában kisebb szótárral és megbízhatósággal rendelkeznek. A beszélőfüggő rendszerek általában beszélőadaptívak is egyben, azaz használatukhoz szükséges egy előzetes betanítási fázis, ezután azonban több szót és jobb megbízhatósággal képesek felismerni, izolált szavak helyett akár kapcsoltszavas vagy akár diktáló üzemmódban is.

b) Betanításnál milyen típusú adatbázis kell az egyik és a másik rendszerhez?

Beszélőfüggetlen rendszer esetén több beszélőtől szükséges hanganyag, hogy ebből közös jellemző vonásokat tudjunk kivonni a betanítás során a minél robusztusabb működéshez. Beszélőfüggő rendszer esetében pedig a hangok paraméterbecslésére nincs szükség (vagy jóval kisebb adatbázis is elegendő), hiszena betanítási fázis során pont ezeket a paramétereket hangoljuk az adott beszélő alapján. Minden más vonatkozásban

(szótár felépítése, nyelvi modellek stb) a két megoldás nem különbözik, illetve max. a szavak számában.

c) Milyen egyéb szempontokat kell figyelembe venni?

Szótárméret, tematika, a hangkörnyezet (zajos utca v csendes iroda), beszédmodor (spontán vagy dialógusszerű), stbstb.

2007.06.15. vizsga

1.

a) Milyen beszédkódolási eljárásokat ismer?

PCM: Pulse Code Modulation (logaritmikus), ezen belül van az A-law (EU) és -law (USA). Lineáris kvantálás. LPC - lineáris predikció. MPEG (layer 3).

b) Milyen mintaillesztési eljárásokat ismer?

Szabályalapú, statisztikai alapú (HMM - Hidden Markov Model és ANN - Artificial Neuro Network) illetve sablon alapú (DTW).

c) Milyen területeken használhatóak a beszédminősítő eljárások?

Mindenhol, ahol hang- ill. beszédátvitel történik, így mobilhálózat, telefonvonal, VoIP, stb.

d) Mi a megfigyelési valószínűség?

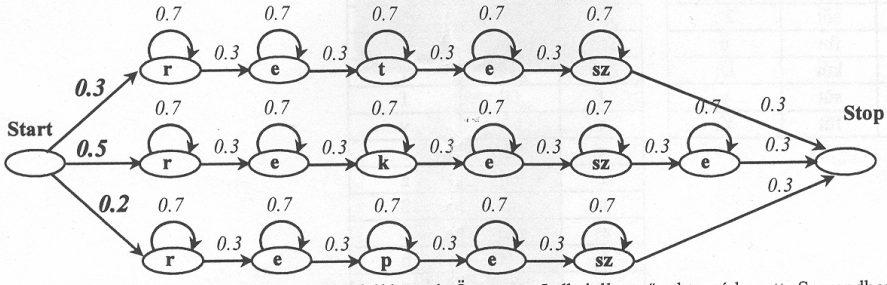
Azt az értéket adja meg, hogy mennyi annak a valószínűsége, hogy egy HMM rendszer x állapotában j jellemzővektort figyeljünk meg.

e) Mi a különbség az akusztikus és a nyelvi modell között?

Az akusztikus modell az egyes beszédhangokra ad egy referencia-jellemzővektorokat, míg a nyelvi modell a beszédhangok kombinációs lehetőségeit adja meg szótárak segítségével, illetve akár a ragozáshoz nyújt megfelelő szabálybázist.

2. HMM

2.



Beszédfelismerési kísérletet végeztünk a fenti HMM hálózattal. Összesen 5 db jellemzővektor érkezett. Sorrendben a 3. vektor megfigyelésének valószínűségei az egyes beszédhangmodellek esetén

$b_{"r"}(o_3) = 0.1, \quad b_{"e"}(o_3) = 0.6, \quad b_{"t"}(o_3) = 0.2, \quad b_{"k"}(o_3) = 0.8, \quad b_{"p"}(o_3) = 0.8.$

El tudjuk-e dönteni ez alapján, hogy mi a felismert szó? Ha igen, mi volt az, ha nem, miért nem? (15 *pont*)

El tudjuk dönteni. Mivel HMM-ről van szó, és a mintaillesztéshez feltétel hogy a START állapotból a STOP állapotba jussunk el úgy, hogy eközben lépések (állapotváltások) és megfigyelések váltogassák egymást, könnyen látható hogy a középső szó (rekesze) kiesik, hiszen 6 állapotot tartalmaz, míg nekünk 5 megfigyelési vektorunk van, így ezen az úton nem juthatunk el a STOPig. Másrészt megfigyelhető hogy a rekesz ill. repesz szónál is minden állapotváltás valószínűsége rendre megegyezik, sőt egyetlen állapotban, a középsőben különböznek (k vs p). Ebből triviálisan adódik hogy az egyetlen különbséget a két út valószínűsége között az adja, hogy mekkora a kérdéses középső állapotban a 3.

jellemzővektor megfigyelése, minden más valószínűségi szorzótényezőben (állapotváltások és megfigyelések: mindig rendre ugyanazt kell megfigyelni ugyanabban az állapotban) megegyeznek.

Mivel p állapotban O3 megfigyelése 0.8, és t állapotban csak 0.2, a "repesz" szó lesz a felismert szó.

És mivel az elején sem egyformák a valószínűségek, azt is bele kéne venni... 0.3*0.2 vs. 0.2*0.8 de így is repesz. -- Csádám - 2010.12.14.

3.

a) Mit jelent egy beszédadatbázis szöveganyagának annotálása, és mit jelent a szegmentálása?

Annotálás: címkézés, azaz a megfelelően szegmentált időintervallumokat ellátjuk a megfelelő magyarázatokkal: milyen hangról van szó, hangsúlyos-e, zöngés-e, stb. A szegmentálás pedig a hanganyag időfüggvényén a hanghatárok bejelölését jelenti.

b) Készítse el az alábbi mondat SAMPA fonotipikus átiratát: *„Elmondtam Havadtői Csillának. Odahívta közben azt a csöppséget, aki megfogta a kilyukadt zacskót”* (segédlet a hátlapon)

Nincs táblázatom, ezért a lényeg: ennél a feladatnál a különböző hangváltásokra kell odafigyelni (hasonulások, összeolvadások, rövidülések és kivetések). Ennek a szövegnek esetében konkrétan:

- elmondtam --> elmontam, a d hang kiesik!
- havadtői --> havattői, részleges hasonulás, zöngétlenesedés.
- közben --> köszben, részleges hasonulás, zöngétlenesedés. (kétségeim vannak, "hasonulás" helyett éppen hogy különbözővé válna)
- azt --> aszt, részleges hasonulás, zöngétlenesedés.
- csöppség --> csöpség, rövidülés
- megfogta --> megfokta, részleges hasonulás, zöngétlenesedés.
- kilyukadt --> kilyukatt, részleges hasonulás, zöngétlenesedés.
- másegyéb?
- odahívta --> odahífta: részleges has., zöngétlenedés

(Írásban nem jelölt) teljes hasonulásra példa: anyja --> annya, hagyja --> haggya másik irányban működő: község --> kösség, tizennyolc

4. A hangszalagrezgést elektroglottográf segítségével (10KHz-es, 16 bites lineáris mintavételezéssel) rögzítjük, majd visszajátsszuk. A beszélő a következő szöveget mondta: *"Eljössz velem? Nem megyek. Nem? Bárcsak eljönnél, úgy szeretném!"*
(Volt ZH kérdés is - 2009 ősz)

a)Milyen beszédjellemzőket lehet meghallani egy ilyen hangszalagrezgésről készített hangfelvételből ami biztosan az elhangzott beszédhez tartozik? Legalább hármat soroljon fel.

A következő beszédjellemzőket lehet meghallani: a beszélő neme (F0 frekvenciájából). A mondatok típusa nagyjából (prozódiából, azaz alaphang-változásokból kifolyólag). Ugyanebből kitalálhatók a hangsúlyok helyei is. Beszéddallam. Emellett a zöngés /

zöngétlen hangok határait is nagyjából el lehet találni. Gond a CC és VV kapcsolatoknál van.

b) Hallható-e a beszéd szegmentális elemei közül valamelyik? Ha igen, akkor melyik(ek). Ha nem, akkor miért nem?

Szegmentális szint: a hangok specifikus időtartamai nagyjából kiolvashatók (?), de nem konkrét hang(kapcsolatok)ra, hanem csak általánosan.

c) Hallható-e a beszéd szupraszegmentális elemei közül valamelyik? Ha igen, akkor melyik(ek). Ha nem, akkor miért nem?

Szupraszegmentális szinten: beszéddallam, hangsúlyok, esetleg ritmus, tempó.

d) Lejegyezhető-e a beszélő személy által mondott szöveg?

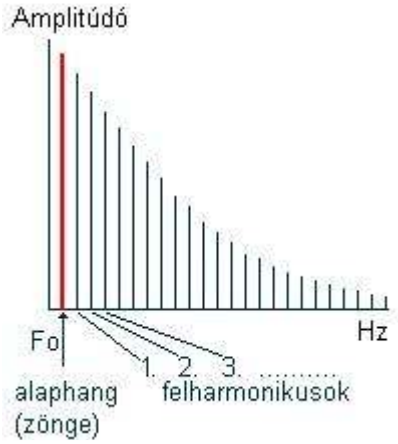
Nem. Rengeteg információ hiányzik, kb csak annyi állapítható meg hogy magánhangzó vagy mássalhangzót ejt az illető, de még ezek határa is nehezen meghatározható.

e) Megállapítható-e a beszélő személy neme egy ilyen hangfelvételből?

Igen. Az alapfrekvencia megfigyelhető, és ebből következtethetünk a nemére is.

f) Rajzolja le a periodikus hangszalagrezgés spektrális képét.

A hangszalagrezgés képe: van egy alapfrekvencia (x Hz, ahol x 100-300 között van), ami a spektrumban egy vonal. Ennek felharmonikusai, azaz többszörösei (n*x Hz) is megjelennek a spektrumban, de egyre kisebb amplitudóval. A csökkenés -12 dB felharmonikusonként. Lásd a képet:



5. Női hangot digitalizálunk8 kHz, 16 bites lineáris mintavételezéssel.
Az átlapolásmentesítő szűrő hibás, az átviteli karakterisztikája a 4000 Hz-es felső határ helyett már 2000 Hz-től levág 60 dB/oktáv meredekséggel. A bemondott üzenet a következő: *„Nyolcezerötszáz lesz a végösszeg.*”

a) Milyen szöveget fogunk észlelni a helyes rekonstruáló szűrővel ellátott visszaállító kimenetén?

Érthetetlen lesz, hiszen rengeteg fontos frekvencia ill. formáns van a 1000-2000 Hz-es tartományban, pl a magánhangzók második formánsának jórésze bele esik ebbe a tartományba. Valami mély mormogást hallunk, gyanítom. (Egyéb, pontosabb ötlet?)

b) Mennyi lesz a jel/zaj viszonya az így elkészített beszédnek?

Jel/Zaj viszony: SNR=1,74+n*6,02=1,74*16*6,02=98,06

c) Mennyire sérül a beszéd dallama a hibás szűrő miatt?

A beszéd dallama nem sérül, hiszen ezt az alapfrekvencia adja meg (F0), aminek a mozgását a hangterjedelem adja meg. Ez pedig tipikusan 100-400Hz közötti érték, amit a szűrő még átvisz.

6. Egy triádos adatbázisú, hullámforma-összefűzéses szintetizátorral a következő mondatot állítjuk elő: *"Miért 40% a határ?"*. Írja le milyen feldolgozási lépések valósulnak meg a példamondaton, amíg a szövegből a végleges hullámforma előáll!
(Volt ZH kérdés is - 2009 ősz)

- Első lépés: begyűjtés! helyett Graféma->Graféma konverziók, avagy a különféle jelölések feloldása, hogy csak betű legyen az output, mégpedig: "Miért negyven százalék a határ?"
- Graféma->Fonéma konverziók avagy a g és y nem külön g és y hanem "gy". Karakterek helyett beszédhangokat írunk. Ezt valami SAMPA átírással lehetne jól leírni.
- Fonéma->Fonéma konverziók avagy nem negyven-nek ejtjük ezt a szót így, hanem netyven-nek. Hasonulások, összeolvadások, rövidülések, kivetések. Eredmény (SAMPA-ban lenne ildomos írni): Mi(j)ért netyven százalék a határ?
- Mindezekkel párhuzamosan fontos a prozódia mondatszintű, szószintű stb lebontása, relatív megadása. Ugyanígy intenzitással is. Amennyire lehetséges, hangsúlyhatárokat is bejelöljük (pl vessző előtt felmegy).
- Ha mindez megvan, egy adatmátrixot kapunk, melyben a szöveg minden lényeges elemét hangokra lebontva megadtuk, ami a kiejtéshez kell. Ezek főbb vonalakban: frázishatárok, szünetek, hangsúly, időtartam, F0, F0 töréspont, intenzitás. Utóbbi 4-et %-ban célszerű megadni.
- Ezt az adatmátrixot kapja meg a triádos beszédgenerátor.
- A beszédgenerátor veszi a hangkódokat a jelölésnek megfelelően. CVC helyzetbe triádot keres, egyéb helyzetekben pedig diádot.
- Ezek hangosságát, frekvenciaszerkezetét és periódusidejét megváltoztatja a megadott százalékoknak stb. megfelelően.
- A szükséges helyekre megfelelő nagyságú szünetet illeszt be.
- Az egyes elemeket simító algoritmusokkal összefűzi.
- Utolsó lépés: a profit!

2001.04.10 ZH

1. Adja meg a megfelelő mértékegységben annak a 80 Hz-es szinuszos hangnak a hangnyomásszintjét, (érzeti) hangosságszintjét és (érzeti) hangosságát, amelynek effektív hangnyomása 0.02N/m2! (15 pont)
Hangnyomasszint, mas neven akusztikai decibel (ld. jegyzet 2. o.)
L=20*lg(P_eff/(20*10^-6*Pa)) [dB] erzeti hangossagszint: ehhez phon gorbesereg kell (zh-n adnak), most ld. 5. o. ha pl. 60 dB szamoltal ki az elobb a 80Hz szinuszos hangra, akkor megnezed, hogy 80Hz-nel melyik gorbe van a legkozelebb 60 dB-hez. Utana nezd meg, hogy ennek a gorbenek (fuggvenynek) mi az erteke 1 kHz-nel, mondjuk legyen 80 dB. Akkor a valasz 80 phon. erzeti hangossag: veszed az elobb kiszamolt phon erteket, es 40 phon -> 2^0=1 son, 50 phon -> 2^1=2 son, 60 phon -> 2^2=4 son, es igy tovabb. van, ahol nem ilyen szepen viselkedik a son-gorbe, olyankor adnak egy abrat.

2. Vázolja egy 100Hz frekvenciájú szinuszjel és egy ugyanilyen alapfrekvenciájú magánhangzó spektrumának jellemző tulajdonságait és ismertesse az ezzel kapcsolatban tanult fogalmakat! (10 pont)

100 Hz szinuszos jel spektrumanak egy nemnulla komponense lesz, pontosan 100Hz-nel, es az erteke a jel amplitudoja (itt nincs megadva az amplitudo). persze a spektrum szimmetrikus az y tengelyre, tehat -100Hz-nel is oda kell biggyeszteni. maganhangzoknal ugye van egy kvaziperiodikus "alapjel" gerjesztes, ami a hangszalaktol jon felfele. ez ffiaknal ~100Hz, noknel ~200Hz, gyerekeknel ~300Hz (ez az f_0). a maganhangzokat ezen kivul azert szeretjuk, mert a spektrumuk formans strukturat mutat. a formansok a spektrumra illesztet burkologorbe maximumai, es f_0 tobbszorosenel vannak. konkret peldak pl. jegyzet 12. o. ide lehet meg irni, hogy a jo megerteshez a komm. eszkoznek at kell vinnie az elso ketto-harom formanst, amit a telefon meg is tesz, ezert a mgh. jol ertjuk telefonban. massalhangzonal (ez itt nem kerdes asszem) ugye a gerjesztes inkabb feherzaj szeru, tehat mindenfele frekvenciakomponens elofordul, es egesz magas frekvenciakon is vannak fontos komponensek, ezeket a tel. nem viszi at, ezert pl. "f" "s" (asszem) nehez megkulonboztetni.

3. Egy 8 kHz-es mintavételi frekvenciával és az alábbi, H(f) karakterisztikájú visszaállítóval működő mintavételező rendszer bemenetére két szinuszos jel összege kerül (jellemzőik: 2kHz, 6Vpp és 5kHz, 2Vpp). H(f) = 1, ha abs(f) ⩽ 3.5; (4 - abs(f)) / 0.5 , ha 3.5 < abs(f) < 4 ; 0, egyébként (a frekvencia mértékegysége [kHz])

a) Milyen jel kerül visszaállításra? (5 pont)

b) Javasoljon egy olyan mintavételi frekvenciát és összetett simító karakterisztikát, amely a fenti jelet helyesen és elfogadható komplexitással megvalósítva átviszi! (15 pont)

Vpp = 2*amplitudo, tehat a 2kHz jelnek 3 [mertekegyseg?] az amplitudoja, az 5kHz-esnek meg 1. mintavetelezed oket 8 kHz, mi lesz a mintavetelezett jel spektruma. szabaly: X kHz mintaveteleztel, a jeled Y kHz volt es A amplitudoju, akkor (n*X)-Y, (n*X)+Y is bejon egy A amplitudoju jel (es ezt minden jelre), n=0,+-1,+-2... azaz frekvencia aplitudo (0*8)-2=-2 3 (0*8)+2=2 3 (0*8)-5=-5 1 (0*8)+5=5 1 (1*8)-2=6 3 (1*8)+2=10 3 (1*8)-5=3 1 (1*8)+5=13 1 (-1*8)-2=.. 3 (-1*8)+2=.. 3 (-1*8)-5=.. 1 (-1*8)+5=.. 1 a minusz szorzok miatt szepen tukros lesz a spektrumod. namost ezt persze nem kell a vegtelensegig szamolni, mert a visszaallito kimeneten van egy szuro, aminek meg van adva az atviteli karakterisztikaja: H(f). namost, hogy megkapd a kimeno jel spektrumat, szorozd be a jeledet a szuro atv. kar-javal, H(f)-fel. tehat ahol H(f) nulla, az kiesik (azokat kiszuri). ahol 1, azokat a komponenseket egy-az-egyben atengedi. es van kozte egy kis atmenet (ez a 4-(abs(f)/0.5). tehat a -2, 2, -3 , 3 fr. komponens megmarad, a tobbi kiesik. ez volt a valasz az a) kerdesre, ezt persze abrazolni kell szepen. a b) kerdesre vagy azt mondod, hogy felviszed a mintaveteli frekveciat (f_0) ugy, hogy teljesuljon az f_0>2B egyenlotlenseg, ahol B a savszelesegge a jelnek, ebben az esetben 5 kHz, azaz azt mondod, hogy a mintaveteli fr. legyen 8 helyett 11kHz, vagy pedig a szurot megvaltoztatod ugy, hogy csak a 2 kHz es az 5 kHz komponens-t ne szurje ki (vigyazz, a spektrum mindig szimm.)

4. Egy jelet másodfokú predikciót alkalmazó rendszerrel viszünk át bináris csatornán.

a) Határozza meg a prediktort, ha R11 = R22 = 1, R12 = R21 = R01 = 0.8 és R02 = 0.6! (10 pont)
b) Rajzolja fel a kódoló és a dekódoló részletes felépítését! (5 pont)
c) Hány bites kvantálót kell alkalmazni a 60dB jel-zaj viszony eléréséhez, ha a predikciós nyereség 30dB? (5 pont)
itt egy lin. egyenletrendszert kell megoldani. R = [R_11 R_12 R_21 R_22], W = [w_1 w_2] B = [R_01 R_02] es R*W=B 2 ismeretlen, 2 egyenlet. matlabbal, mert az jo: > R=[1 0.8
0.8 1]; // az egyutthato matrix > B=[0.8
0.6]; // az eredmeny oszlopvektor > RR=inv(R) // ezzel majd balrolszorozzuk

RR = 2.7778 -2.2222 -2.2222 2.7778

> W=RR*B // most szorozzuk balrol

W = // ez az eredmeny

0.8889 -0.1111 > R*W // ellenorzeskeppen

ans =

0.8000 0.6000 // es visszakaptuk B-t. kiraly.

b) rajzolja fel... nem tudom! c) ezt meg azt mondta a bacsi hogy nem tanultuk

5. Egy telefonos információs rendszerben a következő típusú üzenetet kell bemondani: "A telefonszám: xxx ." ahol xxx bármely magyarországi nyilvános vezetékes ill. mobiltelefon szolgáltató előfizetőjének vagy szolgáltatásának száma lehet. Adja meg a fenti üzenet jó minőségű bemondásához reális erőforrások felhasználása mellett szükséges elemeket! (több jó megoldás is lehet) (20 pont)
ehhez sokmindent lehet irni. beszedszintezis, mert arrol van itt szo, lehet pl. artikulacios: ahol a szajureget probaljak szimulalni. ezt a gyakorlatban nem tudjak jol megcsinalni. formans-szintezis: emberi agy ugy tunik formansok alapjan dont, ezeket probaljak eloallitani (pontosabban hasonlo spektrumu jeleket). figyelembe kell venni formans frekv., formans savszelesseget, mindenfelet. eloallitas lehet soros vagy parhuzamos, egyik ilyen hangokra jo, masik mas hangokra. ez benne van a gordos konyvben is (bme k. konyvtar). osszekapcsolasos: itt elore felvett hangokat kapcsolnak ossze, hogy beszedet allitsanak elo. ennel a peldanal persze ilyet erdemes hasznalni (kell irni). eloszor meg kell hatarozni, hogy milyen elemekbol epited ossze a beszedet: szavakbol, szotagokbol, felszotagokbol, fonemakbol vagy difonemakbol (2 fonema). ha szavakbol allitasz ossze, akkor altalaban nagyon sok szot kell felvenned -> sok memoria kell. szotagoknal is meg ez a baj, felszotagoknal is. de itt jo lenne a beszed minosege. a valosagban altalaban difonemakat hasznalnak, ebbol eleg <100, viszont nem olyan jo a minoseg. ennel a peldanal, ahol csak szamokat kell mondani, persze meg a szavaknak is nagyobb egysegeket valasztunk, szamokat, sot hosszu szamokat. bacsi mondta, hogy eloszor azt kell vegiggondolni, milyen tel. szamokat mondunk vissza: rovidek (segelyhivok): 114,

911 korzetszamok: 1, 20, 30 stb. tk 0<x<100 "sima" szam: lehet 6 v. 7 jegyu pl. 451-333 (videken), vagy hosszu 347-0644 tehat akkor azt mondod, hogy fel kell venned ezeket az x szamokat: 0..999 es ezekbol jonehanyat (a haromjegyueket mindenkeppen, tobbit vegig kene gondolni) ketszer kell folvenni, egyszer ugy, hogy a vegen a hangsuly lemegy, egyszer ugy, hogy felmegy. attol fuggoen, hogy a szam utan mondunk -e meg valamit, attol fugg, hogy melyiket jatszuk le. ide meg sok minden szepet lehetne irni.

2002.04.11 ZH

1. Egyszerre szól egyenként 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500 és 5000 Hz alapfrekvenciájú és 70dB intenzitásszintû hang. Mekkora az ezen komponensekbõl álló komplex hangnak az össz intenzitásszintje? (*10 pont*)

L = 20*lg(P / 20 * 10ˇ-6) = 10*lg(I / 10ˇ-12W)
70dB = 10*lg(I / 10ˇ-12W)
7dB = lg(I / 10ˇ-12W)
10ˇ7 dB = (I / 10ˇ-12W)

10 komponensre:
I = 10 * (10ˇ7 / 10ˇ12) * (10 dB / W) <- lehet, hogy nem helyesen van lemásolva a sor :D
L = 10 * lg ((10* 10ˇ7) / 10ˇ-12) = 10 * lg(10ˇ8) dB = 80 dB

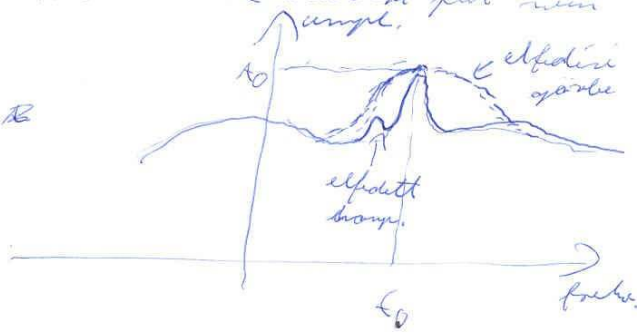
2. Ismertesse a hangelfedési jelenségeket és azok beszédkódolásban történõ alkalmazásának alapelvét! (15 *pont*)

2. feladat

Hangelfedési jelenségek:

- Frekvencia tartományban

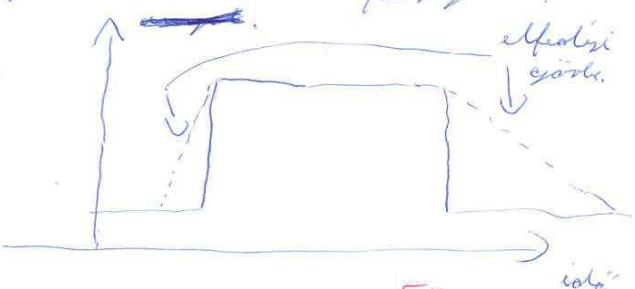
A jelenség lényege, hogy egy nagy amplitúdójú komponens a frekvenciatartományban elfedi a tőle kicsit kisebb frekvenciájú komponenseket, ezeket azokat az emberi fül nem tudja érzékelni.



Egy ilyen hangos komponenshez megadható egy ún. elfedési görbe. Az ezen görbe alatt lévő komponenseket, mert nem fogunk érzékelni. ✓

- Időtartományban:

A jelenség lényege hasonló a frekihez, csak itt egy hangos rész az őt időben megelőző és követő rendszerbeli hangok érzékelését blokkolja. Hasonló módon definiálhatóak itt is elfedési görbék. A hang előtti elfedési görbe lényegesen meredekebb, mint a hang utáni. A hang előtt kb. 20 ms-ig, utána kb. 150 ms-ig van elfedés.



5P

FORDÍTSUNK!

Beszédkódolásban alkalmazás:

Amennyiben egy erős hangot lehódolunk, az elfolási görbéje alatt bármi lehet az elhúzott hangmintákon, úgy a rajzunkat csökkentjük, míg a zaj az elfolási görbe alatt marad. VV

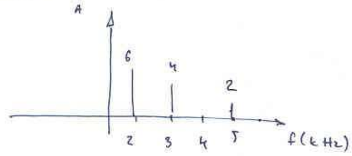
SP

Σ15p

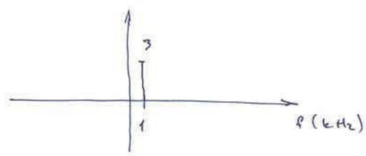
3. Egy ismeretlen mintavételi frekvenciá val és PAM típusú, H(f) karakterisztikájú simító visszaállítóval mûködõ mintavételezõ rendszer bemenetére három szinuszos jel összege kerül (jellemzõik: 2kHz, 6Vpp, 3 kHz, 4Vpp és 5kHz, 2Vpp). A kimeneten egyetlen 1kHz-es komponens jelenik meg 3Vpp szinten.
a) Adja meg a rendszer ismeretlen jellemzõit, amelyek mellett ez a kimenet elõállhatott (több jó megoldás is lehet!!!)! (*10 pont*)
b) Javasoljon egy olyan mintavételi frekvenciát és összetett simító karakterisztikát, amely az eredeti jelet helyesen és elfogadható komplexitással megvalósítva átviszi! (15 *pont*)

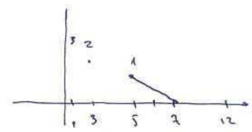
③

Bemenet:



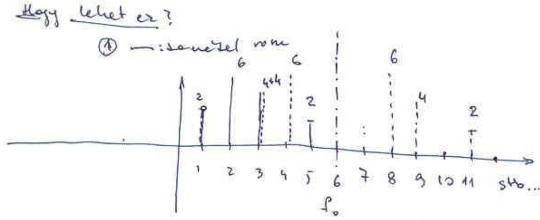
Kimenet:





a) Hogy lehet ez?

① mintavétel van



Magyarázat: önmagukba "tükröződnek" a jelek (a mintavételi frekvencián túl kívül), és pl. az $f_0 = 6$ (kHz) körül megjelenő jel −5 kHz-es komponense az 1 kHz-be került!, ezt még erősíteni kellett 1,5-szeresre, tehát a H(f) az 1 kHz-nél 1,5, 2-től fölfele pedig nulla (ennyi biztos!)

② szűrés:

12?

$$H(f) = \begin{cases} 1{,}5, \text{ ha } f = 1 \text{ kHz} \\ \text{TELJESEN MINDEGY, ha } f \in (1;2) \cup (0;1) \\ 0, \text{ ha } f \geq 2 \text{ kHz} \end{cases}$$

10 p

b) Mintavételezzünk például 12 kHz-en, a szűrés pedig legyen:

6 jobb lenne

$$H(f) = \begin{cases} 1, \text{ ha } f \in (0;5] \text{ kHz} \\ -0{,}5|f| + 3{,}5, \text{ ha } f \in (5;7) \text{ kHz} \\ 0 \text{ egyébként.} \end{cases}$$

ez így jó lesz... (de csak azért, mert nincs 5,5 kHz-es jel!)

14 p

4. Egy telefonos információs rendszerben a következõ típusú üzenetet kell bemondani: "A *gépkocsi rendszáma: xxx* ." ahol *xxx* bármely magyarországon legálisan üzembe helyezett gépkocsi rendszáma lehet. Adja meg a fenti üzenet jó minõségû bemondásához reális erõforrások felhasználása mellett szükséges elemeket! (több jó megoldás is lehet!!!) (*20 pont*)

(4)

- kötött szókészletű beszédinformációs rendszer
  telefonos bemondás
- tematika : gépkocsi rendszám
- a bemondandó szöveg :

  „A gépkocsi rendszáma : XXX" ✓ XXX : rendszám

  ~~A rendszám formátuma: betű betű betű – szám szám szám betű betű betű betű – szám szám~~ - -

- szótárkészlet
  3 szó : gépkocsi , rendszáma , kötőjel

Betűk: a rendszám bemondásakor :
mivel a rendszámok leginkább nem értelmes karaktersorozatok, ezért betűzve érdemes bemondani, illetve a félreértések elkerülése érdekében is.

Az angol ábécé elemei : 26 db betű ✓ ✓ Betűzés hogyan?

ezeket elég egyszer ~~[illegible]~~ felvenni, mert mindig felvitt hangmagassággal ejtjük, mert sosem szerepel a rendszám végén.

Számok:
rendszámok formátuma : 3 betű – 3 szám
4 betű – 2 szám ✓
5 betű – 1 szám
1 betű – 1 betű 4 szám (különleges rendszámok C, V, P, X, E -vel kezdődőek)

1 szám : értelemszerűen 10 db
2 szám : a könnyebb érthetőség kedvéért 10 – 99 -ig : 90 db ✓
1 – 9 -ig az előzőeket használjuk ✓

3 szám : ha takarékoskodunk, megoldható 1 szám 2 szám formában is (de felvehetjük külön, ekkor 100 – 999 -ig +900 db elemet kell felvenni) ✓

4 szám : nem érdemes külön felvenni, kiforratlanság és érthetőség szempontjából 2 szám 2 szám formában jobb. ✓

Viszont minden számot kétszer kell felvenni hangmagasság felemeléssel és hangmagasság leejtéssel. ~~[illegible]~~

összesen 3 + 26 + 2·(10 + 90) = 229 elem ✓

Vagy ha bemondjuk a 3 gyengébbet is, akkor:

3 + 26 + 2 · 90 + 10 + 900 = 1119 elem

ezeket csak: 0-t felírt hangjelzéssel, 1-9 leírt —II—

ismeret leírt hangjelzéssel

még ez sem túlságosan bonyolult.

1

20

5. Mi lehet az oka annak, hogy egy német nyelvû, 50 hangot tartalmazó diádos adatbázis 4 különbözõ változatban, 7 Mbyte, 3.5 Mbyte, 2.54 Mbyte és 1.27 Mbyte méretben is elkészült? (*10 pont*)

⑤

$\frac{350}{127} \sim \frac{11}{4}$ $\frac{2,54}{1,27} = 2$

A hangminőség a különbség.

telefon vonás { 1,27 Mb: 8 bites kvantálás, 8kHz mintavételezés
2,54 Mb: 16 bites —"— , 8kHz —"—

digitális vilag, pl. PC { 3,5 Mb: 8 bites —"— , 22,05 kHz —"—
7 Mb: 16 bites —"— , 22,05 kHz —"—

A 16→8 bit konverzió könnyen megoldható, de a 22,05kHz → 8kHz annál nehezebb.

A 8 bit csak kb. 48dB-t tesz ki, így csak tűrhetően reverzhető. Ellenben egy gyenge telefon jóval kisebb lehet.

10 pont

6.

a) Magyarázza meg, hogy mit jelent a diád és a triád hangsorépítõ elem a fizikai valóságban. Ismertesse mindkettõ használatának az elõnyeit és hátrányait. Hogyan lehet kiküszöbölni a hátrányokat? (*5 pont*)

b) Milyen hangtulajdonságok határozzák meg egy diádos hangsorépítõ elem fizikai hosszát? Adjon példákat rajzzal. (*5 pont*)

c) A magyar nyelvre kb. hány diádot kell elkészíteni, hogy szöveget lehessen felolvastatni egy beszédszintetizátorral? (*5 pont*)

d) A triádos koncepciójú elembázisba hány triádos elemet célszerû tervezni (magyar nyelv esetén). A triádos koncepciójú elembázisban milyen a triádok és diádok aránya?

6.) a)

diád: két félhang (az egyik csúcs is lehet)
triád: két félhang által közrefogott magánhangzó hármasa (a félhang csúcs is lehet)

pl. SAPKA
-ck v c|c c|c v-
diád triád diád triád

diád előnyei: – kevés elem lesz egy diád alapú adatbázisban (kb. 1500 a magyarban), kevesebb kell rögzíteni
– emberi hangjelenségből áll
hátrány: – kevés memória kell hozzá
– a magánhangzókat kettévághatja, ami a spektrumban és az intenzitásban törést jelenthet a formánsok mozgása miatt
– nehéz növeszteni

triád előnyei: – az előbb említett törések torzítás eltűnik a magánhangzó mindig teljes marad
– természetesebb hangzás
– könnyebb növeszteni
hátrány: – sok munka a felvétel
– sok memóriát foglal (kb. 10000 elem kell magyar esetén összesen)
– sok szöveget kell beolvastatni
– diádokat és egyéb elemeket is igényel a kész adatbázis

5p.

b) Általában a magánhangzókat kettévágják időben, míg pl. egy felpattanó zönghangnál (pl: t) nincs szükség a jel egész idejére, hiszen nagy zaj szünet. Ezért pl. a ta diád rövidebb lehet, mint az oa. 3p.

– Ha a hangsúlyzókat és a hosszú magánhangzókat is figyelembe vesszük, akkor csak a hosszak tovább módosulhatnak.
– Ezen kívül megfontolandó még magánhangzók esetén az alapfrekvencia.

+2p

jó!

a ta diád

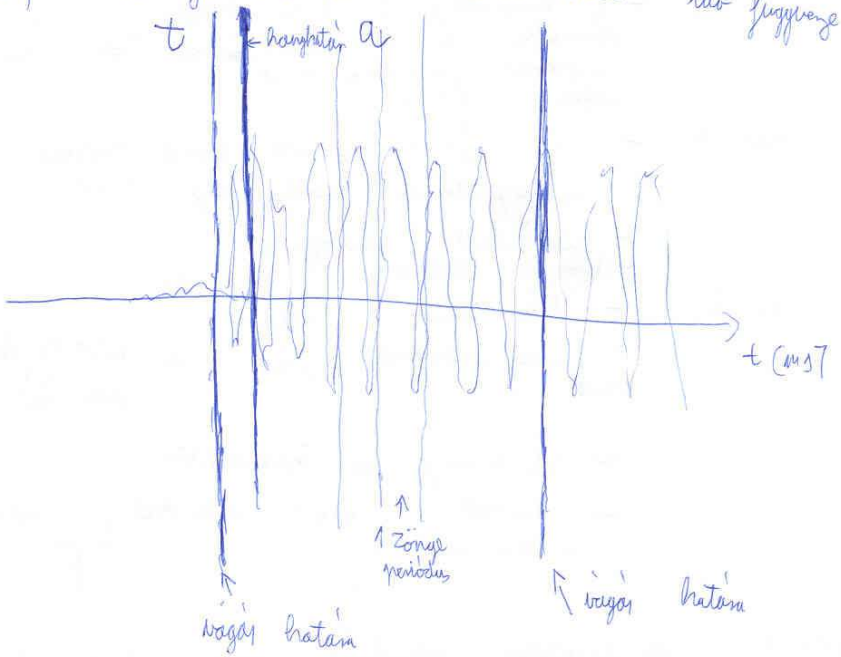
c) kb. 1500 elem. (kb. (40 – mélység) X (40 – mélység))

5p

d) kb. 8000 triád + kb. 2000 egyéb elem.
4 – 5 -ször annyi a triád, mint a diád.

5p

A b) ábrája nélkül: t a diád idő függvénye



2002.05.10. pótZH

2. Mi a pszichoakusztikus modellezés célja? (10 pont)

Pszicho-akusztikus modell:

1: Idő-frekvencia leképzés

- Keretekre vágás: rövid idejű (15-50 ms), átlapolódó (50%) keretek
- Ablakozás
- Fourier transzformáció

2: Pszicho-akusztikus érzeti modellezés

- Az emberi hallás modellezésén alapul, célja a hallható különbségek kiemelése, és a nem észlelhetőek elnyomása
- Monoton legyen a kapcsolat a belső távolság és az MOS között

Pszicho-akusztikus modell elemei:

1: Transzformálás az érzeti tartományra 1a:nemlineáris frekvencia skálák (mel, bark…)

2: Frekvencia elfedés

- Közeli frekvenciák esetén az erősebbik elnyomja a gyengébbet

3: Időbeli elfedés

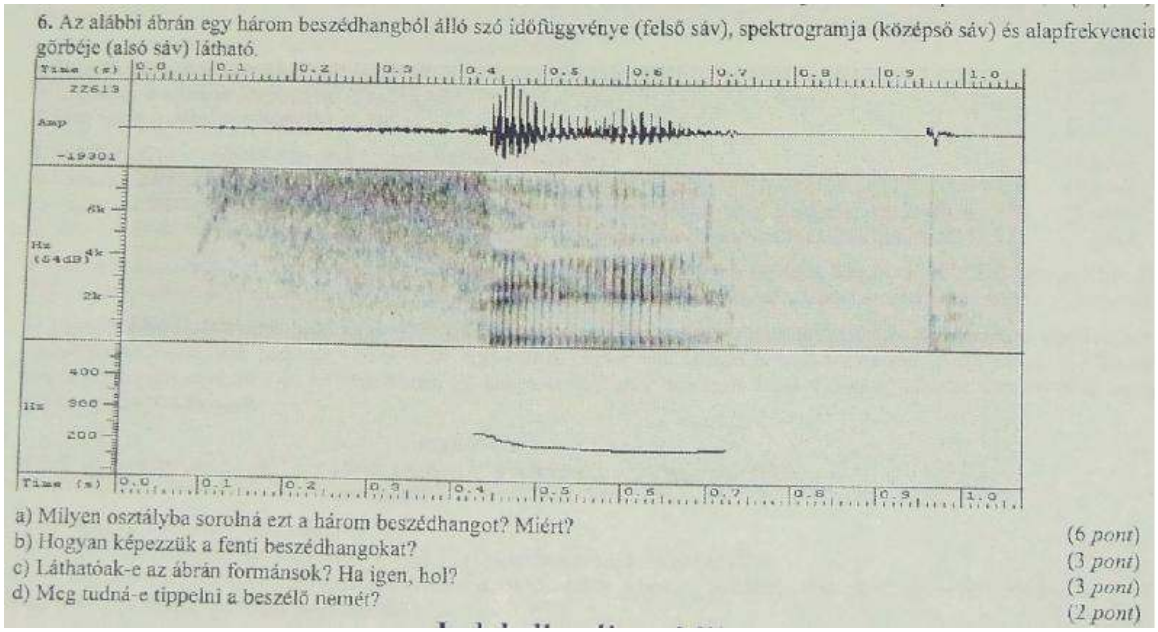
- Egymás utáni rövid impulzusokat egynek hallunk
- Egy erős hang elnyomja a környező gyengébbeket

4: Pszicho-akusztikus hangosság

- Jel energia és hangosság kapcsolata nemlineáris

2007.04.02 ZH

6.



a) 1. hang (0 - 0,43 ms) : fehérzaj szerü (nincs hosszú ü-m) hang, nincs alapfrekvencia, 6kHz feleti komponensei is vannak, nincs néma fázis, nincs zárfelpattanás stbstb, igy (nincs hosszú i-m) tippre: sz, f, v hang lehet. (bővebben: CD) 2. hang (0,43 - 0,73ms) : vannak formánsok (kb. 4 darab erősebb viszszintes vonal a középső sávban), legmagasabban 4kHz körül van. Az alsó sávban látható, hogy nem végig vizszintes, elején van kicsi csökkenés = mikrointonáció, CD alapján meg lehet nézni, milyen mássalhangzó-magánhangzó kombinációknál jelenik meg. Tippre: a,á,ó lehet. 3. hang (0,73 - 1,2 ms) felpattanó zárhang (nem tudom, hogy zöngés vagy zöngétlen, nincs itt a jegyyetem, se a CD, deÖ p,t,k,ty 4-es ből valamelyik. És nem hosszú, mivel a néma fázis 80-90% körüli, hosszú p,t,k,ty esetén ez a fázis hosszabb.

Fót, fát szavakat tippeltem ZH-n.

b)

c) 0,42 - 0,73 másodperc között, ahol igy egy magánhangzó van. 4 darab látható belőlük: a vastagabb vizszintes vonalak.

d) nő, a 200Hz-s alapfrekvenciából lehet erre következtetni. férfiaknak 125Hz, nőknek 200Hz, gyerekeknek 300Hz körüli.