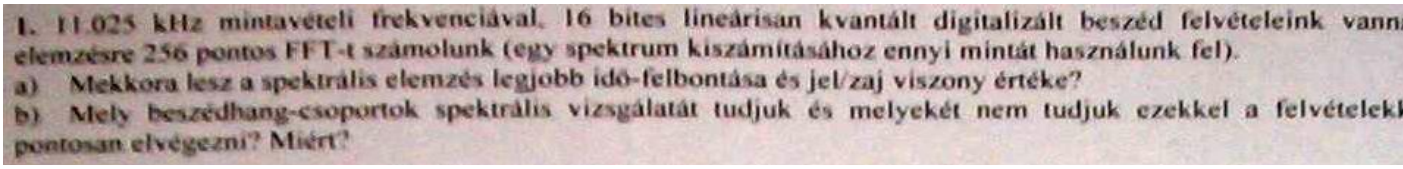
2007 május 25-i vizsga megoldása

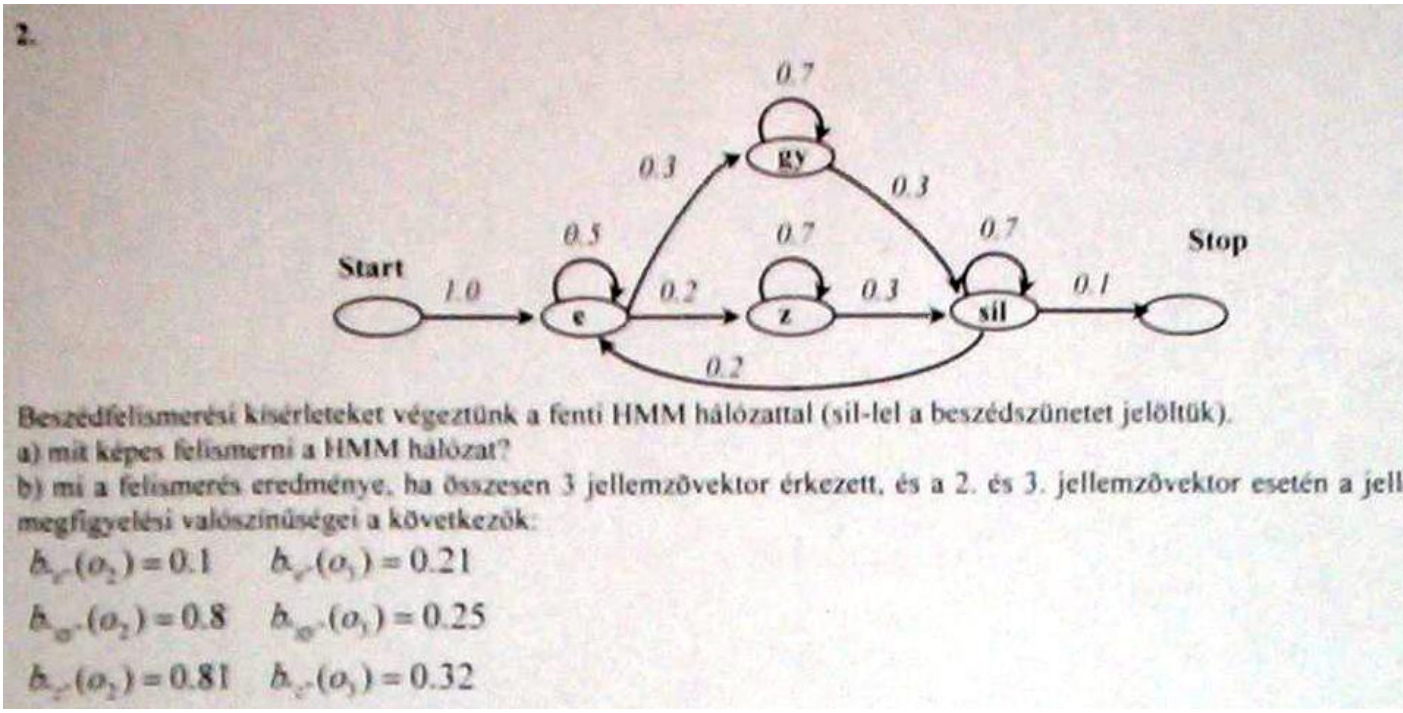
Feladatsor: InfoSite - 2007 május 25. (A Csoport)

1. feladat


1. 11.025 kHz mintavételi frekvenciával, 16 bites lineárisan kvantált digitalizált beszéd felvételeink vann
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?
b) Mely beszédhang-csoportok spektrális vizsgálatát tudjuk és melyekét nem tudjuk ezekkel a felvételek
pontosan elvégezni? Miért?

a. Idő-felbontás: 256 pontos, és 11,025 KHz --> 90,7 μsec, innen az időfelbontás: $256 \cdot 90,7 \mu sec = 23,2 msec$.
   SNR=1,74+n*6,02=1,74+16*6,02=98,06 dB
b. 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. feladat

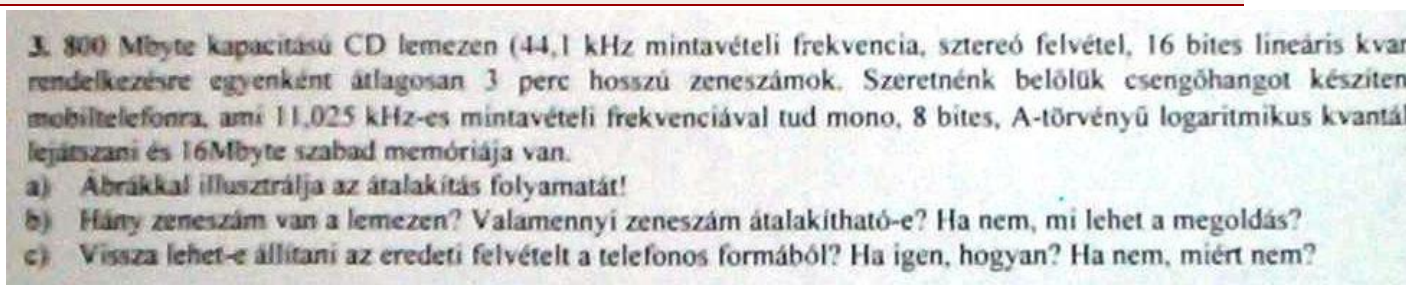


A táblázat nem túl jól olvasható, szerintem az első sorban 'e' utána 'gy' végül 'z' van az alsó indexben.

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 o_2 megfigyelése: 0.81. Átlépés a [sil] állapotba: 0.3. [sil] állapotban o_3 vektor megfigyelése: X. [sil] állapotból STOP állapotba lépés: 0.1.
      Összesen: $0.2 \cdot 0.81 \cdot 0.3 \cdot X \cdot 0.1 = 0.00486 * X$
    - gy[sil]: Átlépés a "gy" állapotba: 0.3. "gy" állapotban o_2 megfigyelése: 0.8. Átlépés a [sil] állapotba: 0.3. [sil] állapotban o_3 vektor megfigyelése: X. [sil] állapotból STOP állapotba lépés: 0.1. Összesen: $0.3 \cdot 0.8 \cdot 0.3 \cdot X \cdot 0.1 = 0.0072 * X$

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

3. feladat

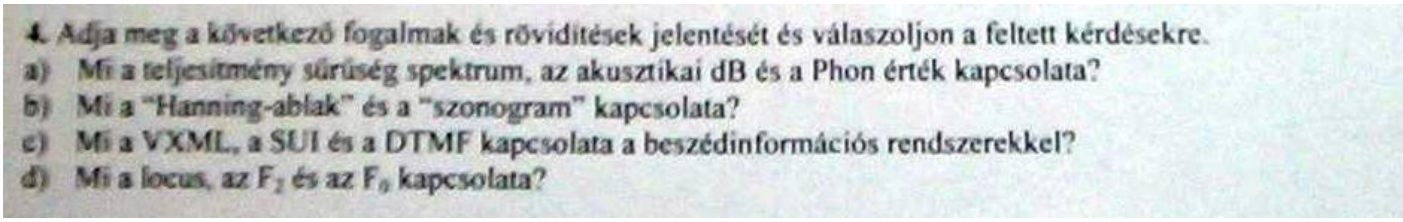

3. 800 Mbyte kapacitású CD lemezen (44,1 kHz mintavételi frekvencia, sztereó felvétel, 16 bites lineáris kvan
rendelkezésre egyenként átlagosan 3 perc hosszú zeneszámok. Szeretnénk belőlük csengőhangot készíten
mobiltelefonra, ami 11,025 kHz-es mintavételi frekvenciával tud mono, 8 bites, A-törvényű logaritmikus kvantál
lejátszani és 16Mbyte szabad memóriája van.
a) Ábrákkal illusztrálja az átalakítás folyamatát!
b) Hány zeneszám van a lemezen? Valamennyi zeneszám átalakítható-e? Ha nem, mi lehet a megoldás?
c) Vissza lehet-e állítani az eredeti felvételt a telefonos formából? Ha igen, hogyan? Ha nem, miért nem?

a. 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. 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. 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. feladat


4. Adja meg a következő fogalmak és rövidítések jelentését és válaszoljon a feltett kérdésekre.
a) Mi a teljesítmény sűrűség spektrum, az akusztikai dB és a Phon érték kapcsolata?
b) Mi a "Hanning-ablak" és a "szonogram" kapcsolata?
c) Mi a VXML, a SUI és a DTMF kapcsolata a beszédinformációs rendszerekkel?
d) Mi a locus, az F_2 és az F_0 kapcsolata?

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.

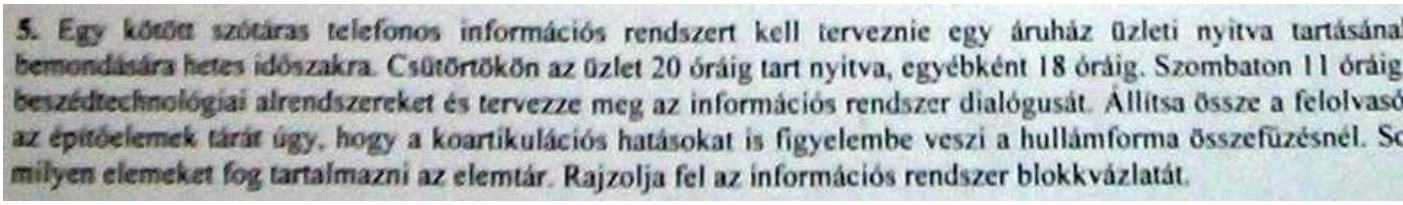
b. Mi a Hanning-ablak és a szonogram kapcsolata?
   - Ha gördülő spektrumot avagy szonogramot szeretnénk készíteni, akkor az időben folytonos jelünket bizonyos kis szeletekben mintavételeznünk kell. A kis kivágott időintervallumokból akkor kapunk jó spektrumot, ha azt megfelelően kiablakozzuk és nem 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. feladat


5. Egy kötött szótáras telefonos információs rendszert kell terveznie egy áruház üzleti nyitva tartásának bemondására hetes időszakra. Csütörtökön az üzlet 20 óráig tart nyitva, egyébként 18 óráig. Szombaton 11 óráig beszédtechnológiai alrendszereket és tervezze meg az információs rendszer dialógusát. Állítsa össze a felolvasó az építőelemek tárát úgy, hogy a koartikulációs hatásokat is figyelembe veszi a hullámforma összefűzésnél. Sc 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.

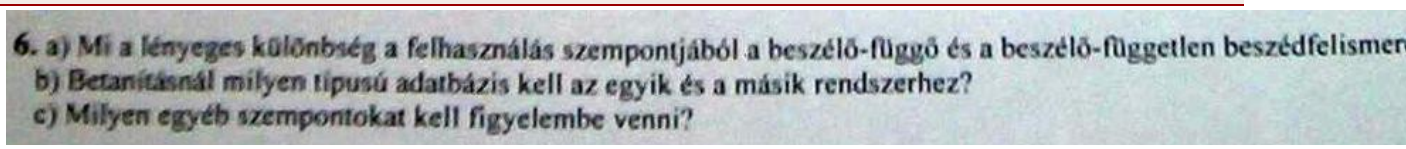
A beszédfelismerő lehetne egy HMM-s rendszer pár szóra (kis szótár) minél robusztusabban (zajra érzéketlen, beszélőfüggetlen) betanítva. A

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. feladat


6. a) Mi a lényeges különbség a felhasználás szempontjából a beszélő-függő és a beszélő-független beszédfelismer
b) Betanításnál milyen típusú adatbázis kell az egyik és a másik rendszerhez?
c) Milyen egyéb szempontokat kell figyelembe venni?

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

-- Gabo - 2008.05.28.

-- Csapszi - 2008.05.29.

-- Maco - 2010.01.06.