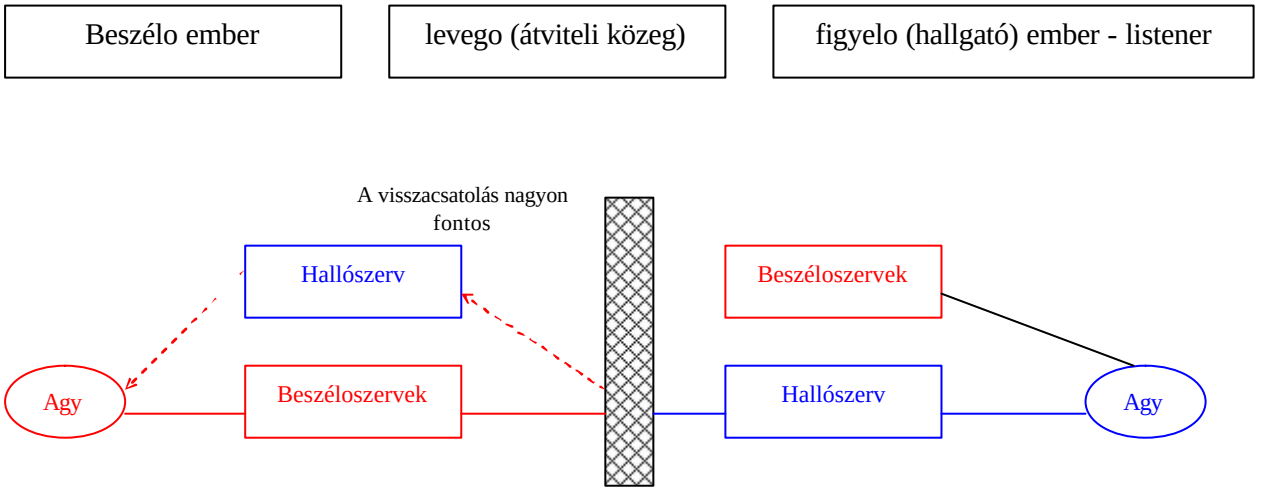
1. BEVEZETÉS

1.1. Alapfogalmak

1.1.1. Nyelv

- Minimum az emberi kommunikáció illetve az emberi gondolkodás legfontosabb eszköze.
- A nyelv elsodleges megnyilvánulási formája a beszéd (a beszéd az emberi kommunikációnak nem az egyetlen formája – nem verbális kommunikáció)
- természetes körülmények között az emberi kommunikáció alapveto jellemzoje a multimodalitás, de a beszéd az egyetlen olyan kommunikációs eszköz, mely önmagában is értheto, ezért a beszédnek az emberi kommunikációban kimelt szerepe van

1.1.2. A természetes beszédlánc



1.1.3. Beszédfeldolgozás (beszédtechnológia)

A *beszédfeldolgozás* a természetes beszédlánc egy vagy több elemének mesterséges (gépi) feldolgozásával foglalkozik

Primer alkalmazások

- *beszédátvitel:* idobeni vagy térbeli távolságon keresztül és mindezt hatékonyan (sávszélesség éhség). A 30-as években került napirendre az az igény, hogy a beszéd sávszélességigényét úgy csökkentsék, hogy a felismerhetosége ne vagy csak alig romoljon. Később ennek módja a digitalizálás és tömörítés (MPEG)
- *beszédszintézis:* a beszéd mesterséges eloállítása. Célja az informatikai folyamatok segítése.
- *beszédfelismerés*
- *beszélo azonosítás:* belépteto rendszereknél alkalmazzák, egy elore bemondott mondat alapján azonosítják a beszélot
  beszélo felismerés: nagy adattárakban beszédmintákat tárolnak, a rendszernek ez alapján kell felismernie, hogy ki a beszélo, vagy esetleg nincs is rá vonatkozó információ az adatbázisban
- beszédkorrektor: pédául beszéd vizualizálása, hogy a süket ember is megtanulhasson beszélni

- beszédmanipuláció: pl. megváltoztatni a beszéd sebességét úgy, hogy a hangmagasság (és hangszín) változatlan maradjon
- ember-gép kapcsolatok megváltozása: a kezelés és szemlélésen keresztül megvalósuló hagyományos ember-gép kapcsolatot felváltja a verbális ember-gép kapcsolat
- beszédinformációs rendszerek: a beszédfeldolgozás az információs rendszerek belso, inherens részévé válik (pl. bemondott telefonszám alapján muködo tudakozó)

1.2. A hang fizikai leírása

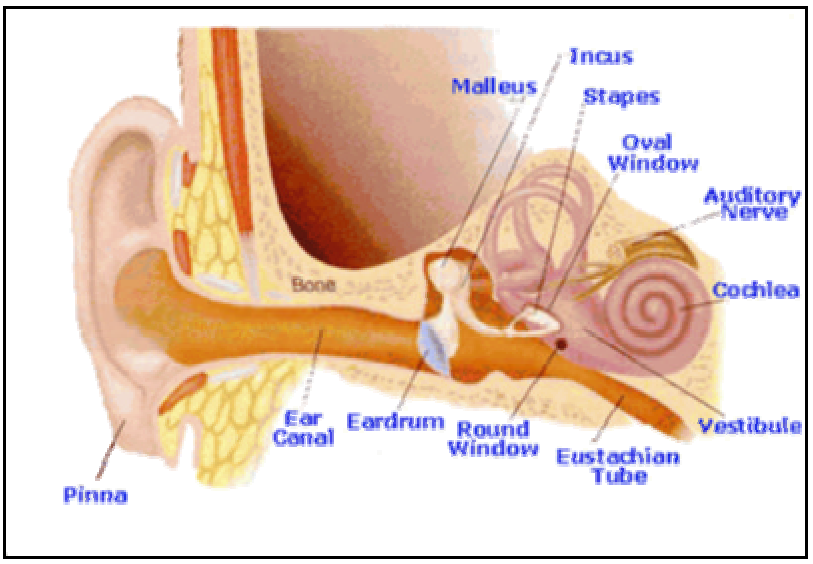
- a hangot le lehet írni nyomással (p) illetve térfogatsebességgel (v)
- P_0 szinten lévo nyomás (1 atm) és ezen picurka longitudinális hullámok
- $P = P_0 + p(t)$
- $P_{eff} = \sqrt{\int (P_0 + p(t)^2 dt}$
- a hang a levego nyomásváltozása, a levegoben longitudinális hullámként terjed (a térfogatrészek mozognak)
- normális viszonyok között: $\frac{p}{v} = 410 \frac{kg}{m^2 s}$, ha síkhullámnak tekintheto

  $c = f\lambda = 340 m/s$
- a hang energiát visz magával (intenzitása van), $[I] = \frac{W}{m^2}$ területegységen idoegység alatt áthaladó energia
- a hangnyomásszint (akkusztikai decibel, intenzitásszint),

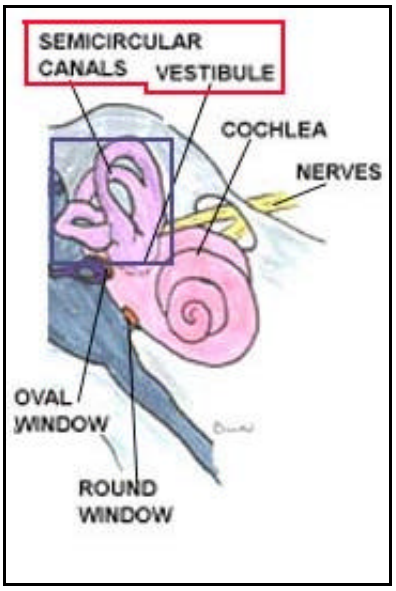
$$L = 20 \cdot \lg \frac{p}{20 \cdot 10^{-6} Pa} dB = 10 \cdot \lg \frac{I}{10^{-12} W} dB$$

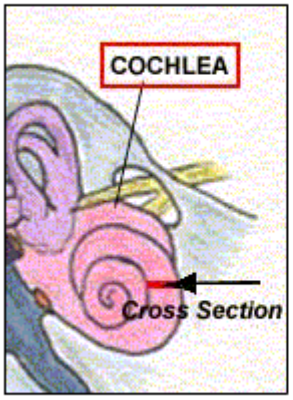
1.3. Hallás

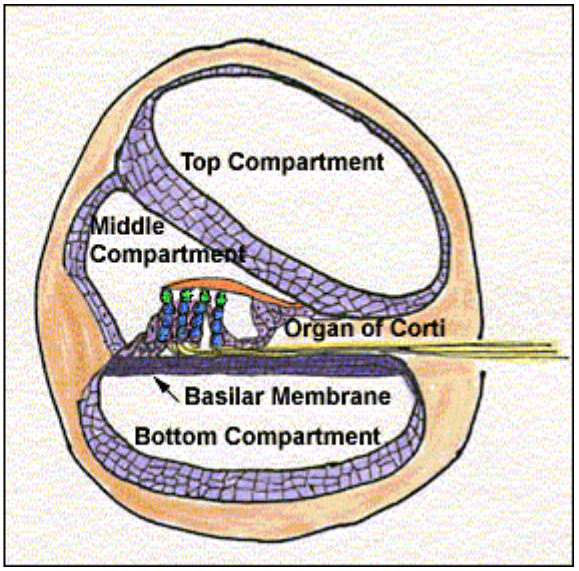
Hallószervek: fülkagyló, külso fülcsatorna (3000Hz rezonanciafrekvencia), dobhártya, halócsontok (kalapács, üllo, kengyel), belso fül



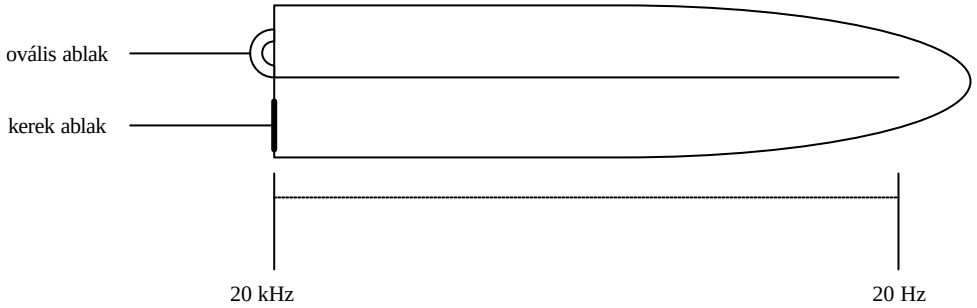
A belso fülhöz a hallócsontokhoz csatlakozik (kengyel, így veszi át a mechanikai hullámokat) egy tömlocske (auditoria tube), amely folyadékkal van kitöltve. Ennek közepén van az alapmembrán, mely hang hatására rezgésbe jön.A membránon 3 sor ún. szorsejt (kb. 30000db) helyzekedik el, ezen sejt végéhez idegek csatlakoznak, melyek közvetlenül az agyba mennek. Egy-egy ideg több ilyen szorsejttel is kapcsolatban lehet. A szorsejtek hozzáérnek a corti-szerv hártyájához, így a mozgás hatására elektromos kisülések keletkeznek, amely az idegpályákon továbbterjed.

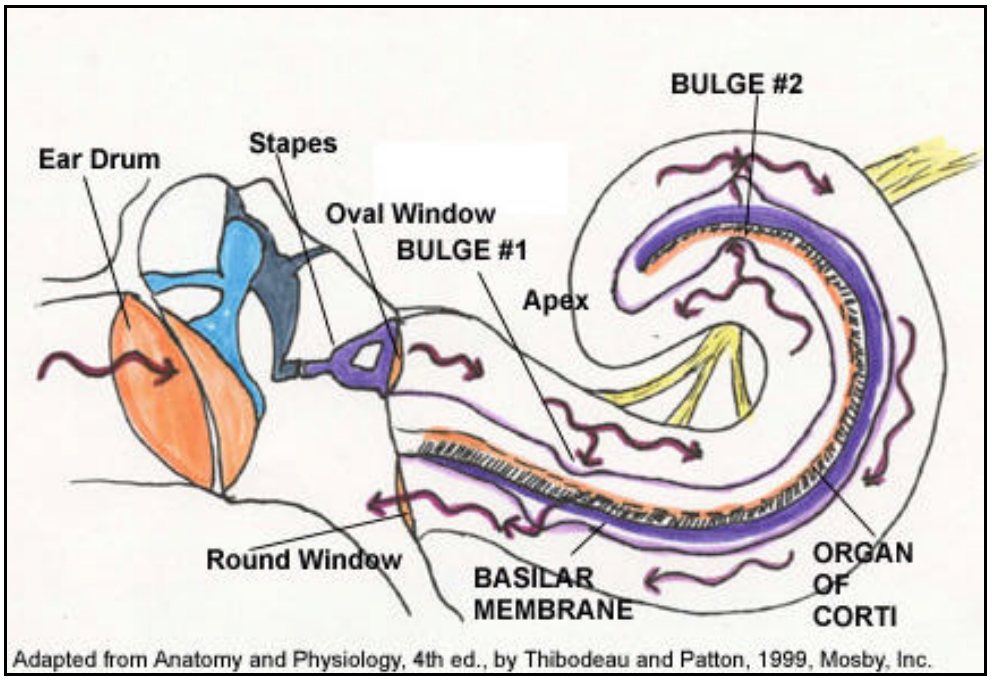






Az elektromos kisülés helye arra jellemzo, hogy milyen frekvenciájú hangot hallottunk. Tiszta szinuszos hangoknál az ovális ablaktól való távoldág számítható. A hang intenziátsát az elektromos impulzusok surusége (frekvenciája) mutatja. A hallás során az agyból is jönnek jelek – gátló jelek. Ezek teszik az ember frekvenciameghatározó képességét ilyen pontossá (a legjobban gerjesztett sejtek környezete blokkolódik).





A hallószervbol jövo idegek nem közvetlennül mennek az agykéreg azon területére, ahol a hallásérzet keletkezik (mint pl. a szemidegeknél), hanem 5 központon keresztül. A hallásmechanizmus a legbonyolultabb az emberi érzékelések között, utnzására egyelore semmi esély.

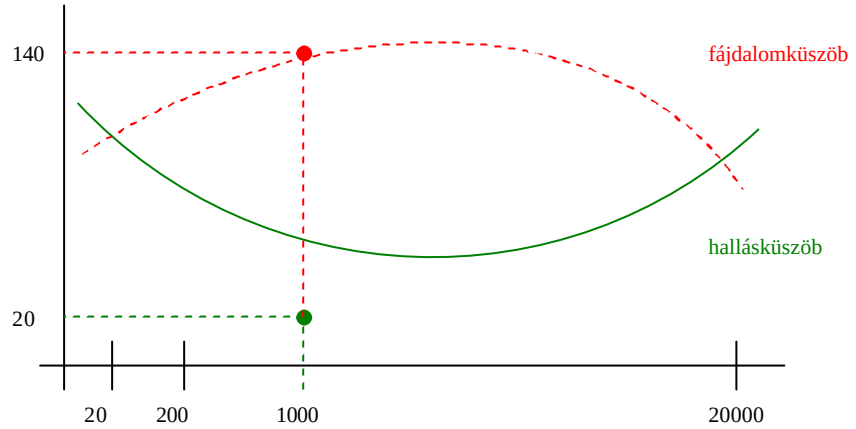
1.4. A hallás pszichofizikai (pszichoakkusztikai) jellemzoi

A hangjelenséggel kapcsolatos mérheto, fizikai mértékek és a hangérzet között nincs 1-1 értelmu megfeleltetés.

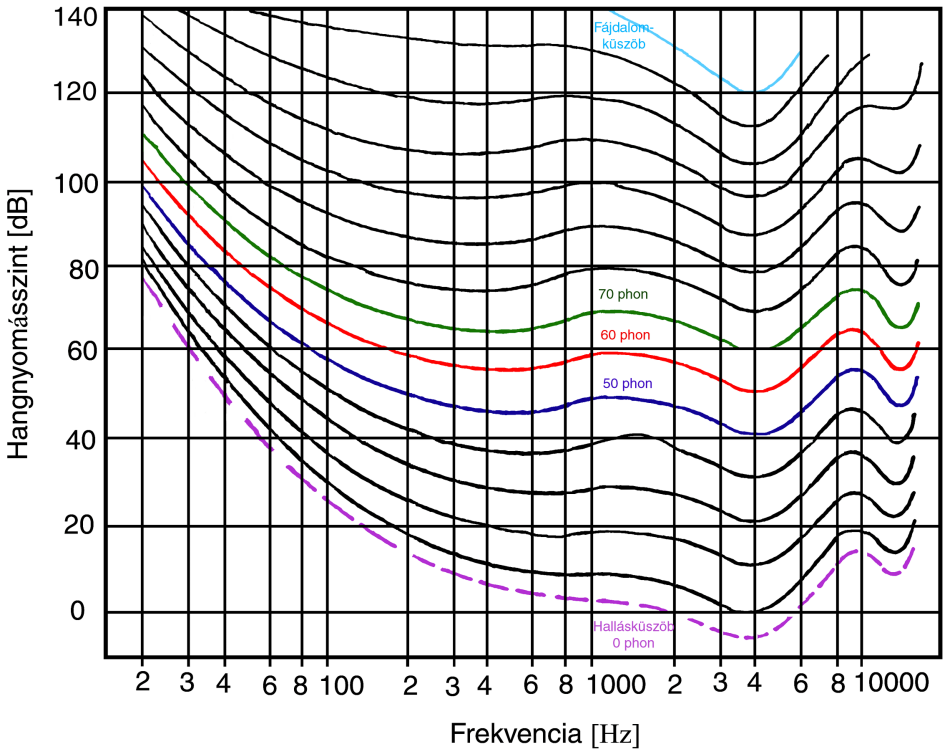
fizikai		hangérzet
intenzitás		hangosságérzet
spektrum		hangszín
frekvencia		hangmagasság

1.4.1. Azonos hangosságú (phon) görbék

Ezeket tiszta, szinuszos nagokra vizsgálják. Az *n-phon* az a görbe, amelyet az emberek statisztikailag azonos eroddégunek hallanak és 1kHz-en *n dB* az erossége.



- mértékegység a [phon]=hangosságszint
- a phon elég jól megadja a hangosságérzetet, de nem teljesen (pl. 30 phon mennyivel kevesebb, mint 40 phon?)



1.4.2. Hangosságérzet

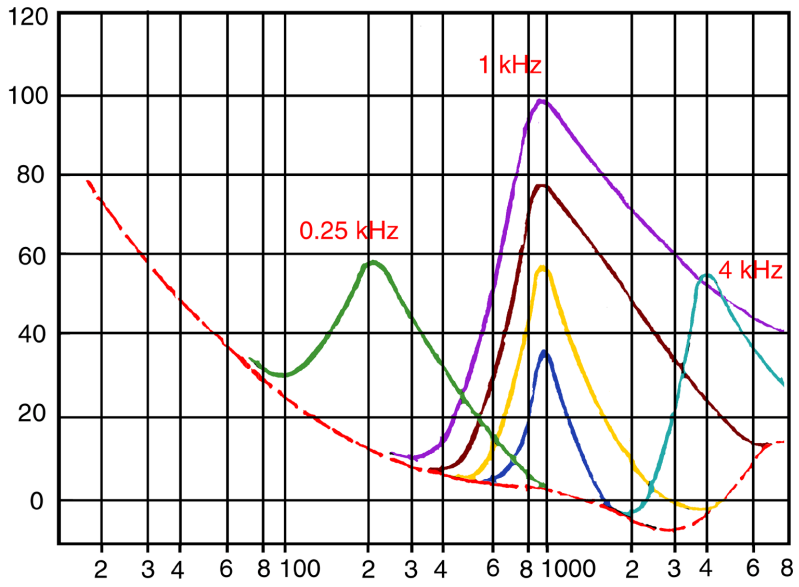
- általában son-ban adják meg
- megmutatja, hogy két phonban mért hangosságérzet aránya mekkor
- egységnek a 40 phon-t tekintik, az 50 phon 2 son, és így tovább

1.4.3. Kritikus sávok

- alkalmazzunk egy keskenysávú fehérzaj-gerjesztést (ennek intenzitása a görbe alatt lévo terület)
- a következo vizsgáló jel legyen szélesebb spektrumú, de ugyanolyan intenzitású
- bizonyos spektrumszélesség fölött a hangosságérzet no annak ellenére, hogy a kibocsátott zaj intenzitása nem változik
- **kritikus sávnak** [1] azt nevezzük, amelyen belül a hangosságérzet nem érzékeny a sávszélességre.
- a kritikus sávokat kísérlettel szokták meghatározni
- bark (zwicker) skála: két fekvencia különbségét pszichoakkusztikai szempontból megmutató skála; a különbség jellemzoje, hogy hányszor lehet felmérni a kritikus sávszélességet

1.4.4. Elfedés a frekvenciatartományban

- elfedo hangjelenség: 1000Hz-en megszólaltatunk egy nagy intenzitású hangot, és a kritikus sávszélességen belül szóló kisebb intenzitású hangokat a fülünk nem érzékeli
- pl. f = 1000 Hz, Δf = 160 Hz, L = 80 dB esetén a 1010Hz-en 40 dB intenzitássaá szóló hangot nem érzékeljük
- tehát ha van elfedo hang, akkor a többi hang számára a hallásküszöb megemelkedik
- ezek az elfedési görbék alacsony frekvenciákon keskenyebbek, magasabb frekvenciákon pedig szélesebben elterülnek
- ezeket a tulajdonságokat hangtöömörítésnél alkalmazzák elsosorban



[1] [1] Kitéro: A spektrálsuruség

Tekintsük véletlenszeru függvénysokaság egy elemét (ilyen módon sokmindent érdemes modellezni, pl. emberi beszédet, zajt, stb.). Ezt a sokaságok sztochasztikus folyamatnak nevezzük. Jelöljük a sztochasztikus folyamatot x_t-vel, ennek egy elemét $x_a(t)$-vel, $s_x(f)$-fel pedig ennek a sztochasztikus folyamatnak a Fourier transzformáltját. A $x_a(t)$-t rávezetve egy keskenysávú szurore (f_a és f_b között) mérjük a folyamat teljesítményét. A $x(t)$-sztochasztikus folyamat spektrálsurusége megmutatja, hogy f_a – f_b tartományban ennek a sztochasztikus folyamatnak mennyi a teljesítménye.

$$P(\{f_a, f_b\}) = \int_{f_a}^{f_b} s_x(f)df$$

1.4.5. Elfedés az idotartományban

- ha van egy nagyobb intenzitású hangjelenség, akkor ha ennél kicsit kisebb intenzitású megszólal, nem vesszük észre
- ha befejezodik a hangjelelenség, a következo megjelenését nem azonnal vesszük észre, a fülnek van egy kis tehetetlensége – ez akár 150ms is lehet
- a fenti jelenség „visszafelé” is muködik, csak sokkal kisebb idoértékkel – 20ms

1.4.6. Irányérzékenység

- kisfrekveciákoin az irányérzékenység a két fülbe érkezo jel közötti fáziskülönbségbol adódik
- nagyobb frekvenciákon az irányérzékenység az intenzitáskülönbségen alapul

1.4.7. Frekvencia – ido felbontóképesség

- Kérdés: milyen hibával (Δf) találjuk el a Δt ideig tartó hang frekvenciáját? $\Delta f := s$
- Harkevics és Gábor Dénes bebizpnyították, hogy lineáris rendszerekben $\Delta f * \Delta t \approx 0{,}01$.
- Ugyanezt a fül kb. 1000 Hz-ig jobban csinálja, noha különbözo maszkolási jelenségekkel becsapható, azért nem annyira, mint a szem (pl. a szem számár RGB-bol gyakorlatilag minden szín kikeverheto)

1.5. A beszéd nyelvi szerkezete

- A hangot kétféle szinten vizsgáljuk:
  - akkusztikai szinten: valamilyen hanghullám
  - agyi szint: képesek vagyunk ezeket valamiféle diszkrét elemek (hangok) sorozatára bontani (ezen diszkrét elemek a beszédhangok, melyeket a hallás során az ember érzékelni képes)
- Lehetséges hierarchiaszintek: beszédhang – szótag – szó – mondatrész – … – mondat – … ezek nem mindegyike precízen meghatározott
- egy diszkrét beszédhang megszámlálhatatlanul végtelen sok alakú idofüggvénybol absztrahálódik
  - intraindividuális: ugyanazon ember ugyanazta a hangot kétszer egymás utaán nem ugyanúgy mondja
  - interindividuális: két különbözo ember egyanazt a nangot nem ugyanúgy ejti ki
- A beszéd (nyelvek) úgy alakult(ak) ki, hogy ritkák az átlapolódások. Nem használja ki az összes lehetséges hullámformát (redundáns)
- **Artikulációs (akkusztikai) bázisnak** nevezzük a beszédkeltés folyamán használt elemi folyamatokat, és **percepciós bázisnak** hívjuk a beszéd megértése folyamán használt elemi folyamatokat.

1.5.1. Fonetikai megfontolások

- fonéma készlet – elemknek olyan minimális számosságú halmaza, amelybol minden közlemény jelentéshelyesen, de csak egyféleképpen állítható elo agyi szinten, vagyis ha egy közleményben egy fonémát kicserélnénk, akkor megváltozna a közlemény értelme vagy elveszíti értelmét.
- allofonok – egyazon fonéma különbözo akkusztikai megjelenései.
  - például a magyarban a nyílt és a zárt *e* megkülönböztetheto (akkusztikailag), de egy fonéma
  - harang szóban az *ng*-t egy hangnak ejtjük, nem külön *n* és *g* egymásutánjaként, de ez a hang nem külön fonéma
  - összeállították a magyar nyelv fonémakészletét, ebben leggyakoribb az *e* (*e*ke) és legritkánn a *h* a*hh*oz).

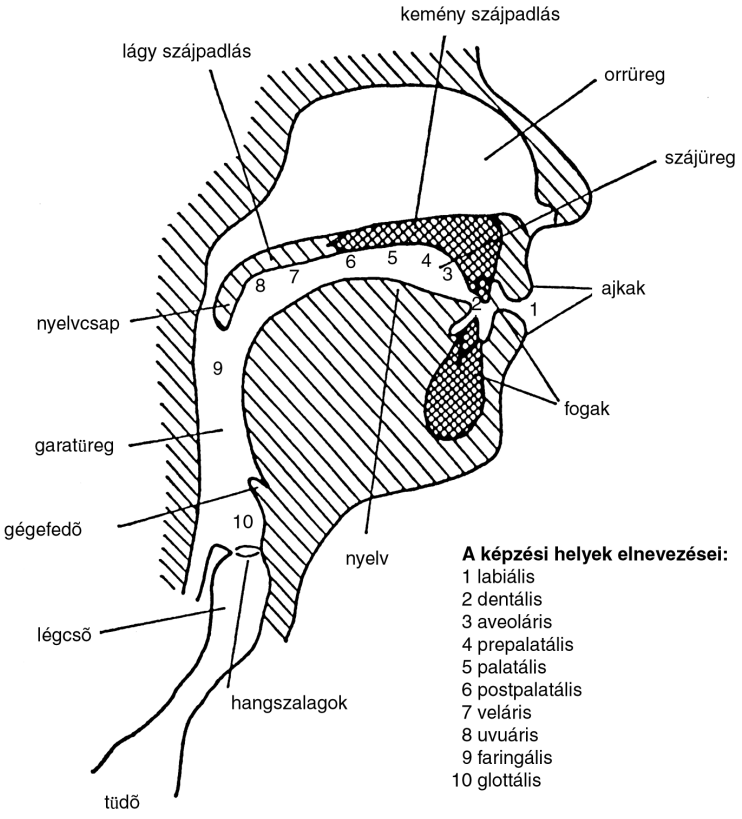
1.5.2. Az írás és a beszéd kapcsolata

- sok nyelvet karakterekkel írnak le, de léteznek olyan karakterek, amelyek szimbólumok (pl. 1, #)
  - ortografikus karakterek: azoka karakterek, amelyeket ki tudunk ejteni
  - graféma: pl. egé*szség* szóban az *sz* hang egy graféma
- szöveg leírásakor ortografikus karaktereket használunk, de ha a jelentésig el akarunk jutni, akkor a graféma-konverziót meg kell tenni. Ezek már fonémákkal írhatók le.
- fonetikus leírások: a hangzást is megpróbáljuk leírni
  - IPA átírás – ASCII karakterekkel nem írhatók le
  - SAMPA – a 7 bites ASCII karakterekkel leírhatók
- a hangok átmenetekkel kapcsolódnak egymáshoz, természetes átkötés (ha nem stimmel, érzékeljük)

1.6. A hangképzés

A *tüdo* energiát szolgáltat, levegot pumpál a *légcso*be. A gégében elhelyezkedo *hangszalagok*at a hangképzés során vagy használjuk, vagy nem. A levego a gégébol a *garatüreg*en keresztül a *száj-* illetve az *orrüreg*be jut. A szájüregben a *nyelv, fogak, ajkak* is részt vehetnek a hangképzésben. A hangszalagoktól felelé sor rész, mely részt vesz a hangképzésben az ún. *vokális traktus* (üregrendszer, toldalékcso).

A hangképzõ szervek vázlatos képe



Alapvetoen három emberi hangkeltési mechanizmus van.[1]

[11] Léteznek más hangképzési mechanizmusok is (pl. csettintés Afrikai törzseknél), de mi nem használjuk oket. Valószínuleg a hangképzési mechanizmusok és maguknak a hangokank is a nyelvben lévo elofordulása attól függ, hogy az adott nyelv milyen körülmények között alakult ki (zajviszonyok, stb.)

1.6.1. Zönge

- a hangszalagok tulajdonképpen két rostos, izmos hártya, melyek összeérnek, az izmok pedig a gégében lévo porcokhoz tapadnak
- ha az izmok megfeszülnek, akkor a levegot nem engedik ki (féldobok)
- a levego kiáralik a tüdobol, és ha a hangszalagokat az izmokkal muködésbe hozzuk, akkor azok elott kis túlnyomás alakul ki (kb. 3-4cm vízoszlop nyomásának megfelelo), melynek hatására a hártya szétfeszül. Emiatt azonban a túlnyomás lecsökken és a hártya ismét visszazár. Így egy kvázi-periodikus jelenség alakul ki. Ennek a kvázi-periodikus jelenségnek a periódusideje férfiak esetében 8-12 ms, noknél 4-7 ms.

1.6.2. Turbulens áramlás

- szignifikáns szukületeket képezünk a vokális traktusban (pl. az *f*, *s* hangok képzésénél)
- a levegorészecskék a szükületet elhagyva véletlenszeruen leszakadoznak
- az, hogy a hang hogyan hangzik attól is függ, hgoy hol a résképzés helye és mi van után

1.6.3. Lökéshullám

- miközben a levegot préseljük ki, a vokális traktusban zárat képezünk (nem a hangszalagokkal) – ezért egy ilyen lökéshullámot 3 szakaszra tudunk bontani
- néma szakasz (néma fázis) – ez a zár képzésének ideje, ilyenkor a levego nem tud továbbhaladni
  - zárfelpattanás – a zár felpattan, és ennek hatására valamilyen hang jön létre (pl. *p* hangnál)
  - elhalás (átmenet)
  - látszólag véletlenszeruen jön, amplitúdóban nagyobb, mint a turbulens áramlás

1.7. A beszédhangok osztályozása

magánhangzók: a, á, e, é ,i, o, ö, u, ü

mássalhangzók

nazálisok: m,n,ng, ny

likvidák: l, j

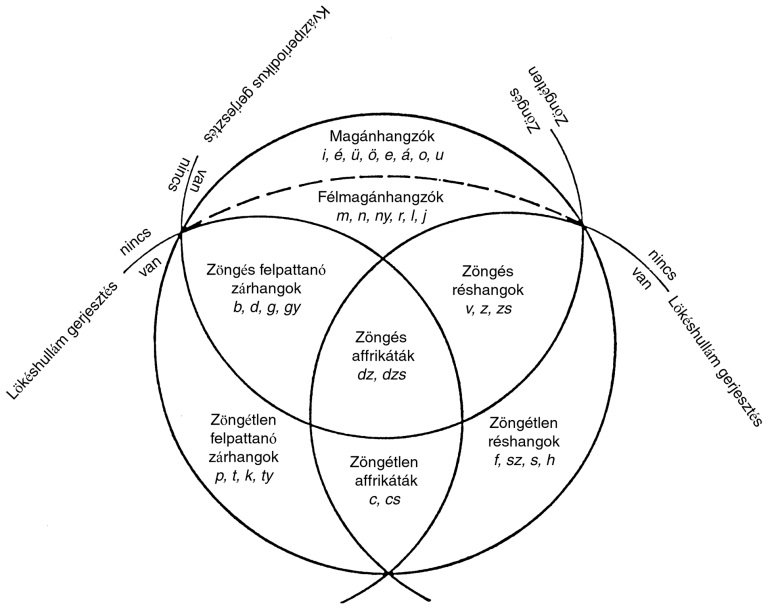
pergo: r

felpattanó zárhangok (explozívák, plozívák, stops): t, d, p, b, k, g, ty, gy

réshangok (frikatívák): f, v, s, zs, sz, z, h

zár-réshangok (affrikáták): c, dz, cs, dzs

A magyar beszédhangok osztályozása a fonetikai és a gépi feldolgozás alapján



1.8. Spektrális módszerek

1.8.1. Periodikus jelek – Fourier sor

$f(t) = f(t + kT_0),\ \forall k \in \mathbb{Z}$

$$f(t) = c_0 + \sum_{n=1}^{\infty} c_n \cos(n\Omega_0 t + f_n), \Omega_0 = \frac{2p}{T_0}$$

- egy jel spektrumán azt értjün, hogy adott frekvencián milyen amplitúdójú és fázisú az adott jel. Tehát a spektrumot a $\{nW_0, c_n, f_n\}$ hármas határozza meg.
- az idoben periodikus jelek vonalas spektrumúak
- a Fourier sor komplex alakja:

$$\cos(x) = \frac{e^{jx} + e^{-jx}}{2} \Rightarrow f(t) = c_0 + \sum_{n=1}^{\infty} \left(\frac{c_n}{2} e^{jf_n} e^{jn\Omega_0 t} + \frac{c_n}{2} e^{-jf_n} e^{-jn\Omega_0 t} \right)$$

bevezetve az alábbi jelöléseket:

$C_0 = c_0,\ C_n = \frac{c_n}{2} e^{jf_n}$ és $C_n^* = C_{-n} = \frac{c_n}{2} e^{-jf_n}$

$$f(t) = \sum_{n=-\infty}^{\infty} C_n e^{jn\Omega_0 t}, \text{ ahol } C_n = \int_{t_1}^{t_1+T_0} f(t) \cdot e^{-jn\Omega_0 t} dt$$

- egy általános vonalas spektrumú jel Fourier-sora

$$f(t) = \sum_{n=-\infty}^{\infty} D_n e^{j\Omega_n t}, \text{ ahol } \frac{\Omega_n}{\Omega_m} \text{ irracionális is lehet.}$$

1.8.2. Egyszeri folyamat – Fourier-integrál

$$F(w) = \int_{-\infty}^{\infty} f(t) e^{-jwt} dt \ = \mathrm{F\{f(t)\}}$$

- Jelentése: kontinum sok szinuszos hullám összege. Ez egy komplex függvény abszolutértékkel és arkusszal.
- az idofüggvényt az inverz-Fourier transzformációval állíthatjuk elo

$$f(t) = \frac{1}{2p} \int_{-\infty}^{\infty} F(w) e^{jwt} dw$$

- fontos paraméter még a jel fajlagos energiája és teljesítménye:

$$E = \int_{-\infty}^{\infty} f^2(t) dt \text{ illetve } P = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} f^2(t) dt$$

- azért fajlagos, mert megfelelo konstanssal szorozva energiát illetve teljesítményt kapunk
- periodikus jelek esetében E = ∞ és P véges, egyszeri folyamatoknál E véges és P = 0.

1.8.3. Lineáris rendszerek hatása az átvitt jelre

- a mínusz végtelentol végtelenig való integrálás probléma, hiszen akkor meg kell várni, amíg a teljes folyamat lejátszódik, erre viszont nincs mindig lehetoségünk
- a teljes folyamatot kiablakozzuk: az idofüggvényt megszorozzuk a t1 idopillanatokra eltolt ablakfüggvénnyel

  $F_w(\omega,t_1) = F\{f(t)*a(t,t_1)\}$, kérdés, hogy mennyire rontja el az ablakolás a spektrumot

  $F_w(\omega,t) = F(\omega)*A(\omega,t_1)$

- legjobban az ún. szeretjük *Hamming-ablakot*, mert ennek spetrumában a fomaximum és a második maximum között 50 dB erosítéskülönbség van, tehát a távoli frekvenciákat a konvolúcióban ez az ablakoás gyakorlatilag nem veszi fig yelembe.
- Hamming-ablak:, $a(t) = 0.54 - 0.46 \cdot \cos 2p \frac{t}{T_a}$
- digitális jelfeldolgozás során DFT-ket (Discrete Fourier Transformt alkalmazunk, általában a DFT-nek egy gyorsan elvégezheto módszerét alkalmazzuk, az ún. FFT-t (Fast Fourier Transform)
- spektrogram: gördülo spektrumot tekintve (ido – frekvencia sík) ahol a spektrum értéke nagy, ott erosen befeketítjük,
  - a spektrális viszonyok az ido függvényében változnak
  - az 1900-as évek közepén az ún. *szonográfot* alkalmazták, amely ezt a görülo spektrumot közelítette (a közelített ábra a *szonogram, spektrogram*)

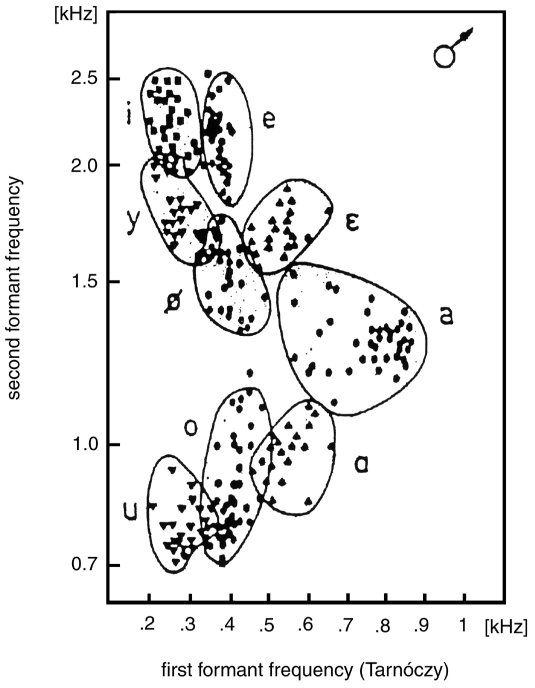
1.9. Beszédhangok finom szerkezete

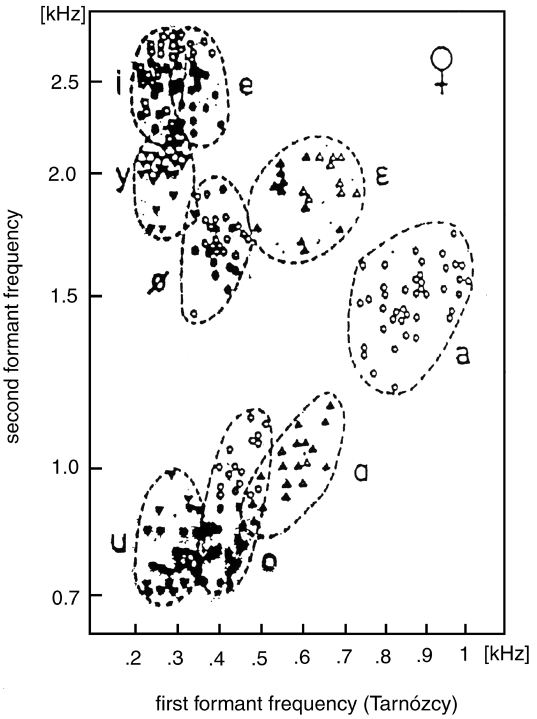
1.9.1. Hosszan tartható zöngés hangok

- olyan hangok, amelyekben a zöngén kívül nincs más hangkeltési mechanizmus
- idotartományban közel háromszög alakú térfogatsebességgel ábrázoljuk
- a hangképzés során keletkezo hang leíró függvénye kvázi-periodikus, ezért Fourier-sorba fejtheto (gazdag felharmnikus tárral rendelkezik)
- pl. a beszédet a telefon 300-3400 Hz között viszi át
- ha az alapharmonikus (f_0) nincs benne az átvitt jelben, a fül akkor is képes azt kiérezni, mert a felharmonikus tár igen gazdag
- a vokális traktuson keresztül a hang a száj illetve orrüregen keresztül távozik. Az orrüreg minden hangra hatást gyakorol direkt illetve indirekt módon. Ezt úgy modellezhetjük, hogy
  - a vokális traktusnak van egy átviteli függvénye: $V(\omega)$
  - létezik egy ún. sugárzás (radiációs) ellenállás: $R(\omega)$, ahogy a szájból eltávozik a hang, a magas frekvenciákat „lenyomja”
  - $R(\omega)*V(\omega)$ a zönge által keltett hangot *formálja*
  - $P(\omega)$ – ezen hang spektrális megjelenése
- a vonalas spektrumra illesztheto burkológörbe maximumhelyeit *formáns*oknak nevezzük (F_1, F_2, … F_n formáns frekvenciák)
- a hosszan tartható zöngés hangok formáns struktúrával rendelkeznek
- hogyan értelmezzük a burkológörbét: minden spektrumvonalra ültessünk egy $\frac{\sin x}{x}$ függvényt. Az adott helyen a függvény értéke legyen a spektrumvonal magassága és a hullámátmenetek F_0 távolságban legyenek (ahol F_0 az alapharmonikus frekvenciája.

1.9.1.1. Magánhangzók csoportja

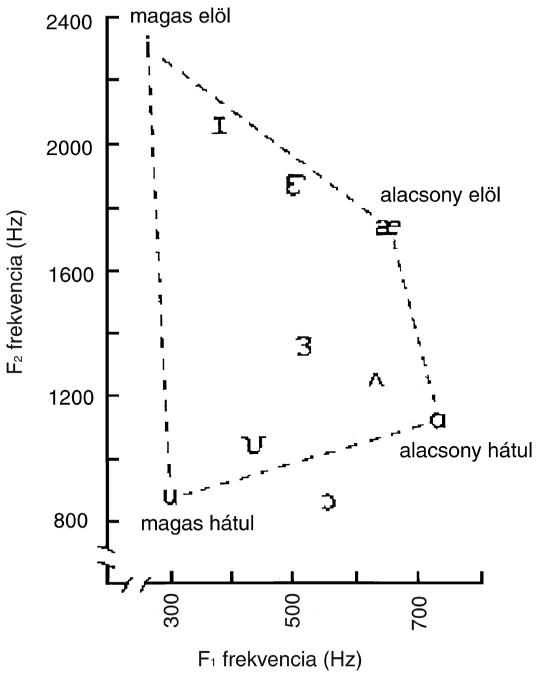
- a magánhangzókat két formáns közel, három formáns szinte teljes biztonsággal megkülönbözteti egymástól, érzeti szempontból leglényegesebbek a formánsfrekvenciák
- az A_1 .. A_n formánsok „amplitúdóit" az elsore (A_1) szokták normálni
- a helyi maximumok alatt 3 dB-lel meghúzott vonal és a burkológörbe metszéspontja jelöli ki B_1, B_2, …B_n formánsok sávszélességét.
- a magánhangzókat F_1-F_2 síkon szoktuk ábrázolni
- a nok F_1-ben és F_2-ben is magasabb területre kerülnek, mint a férfiak, további jellegzetes különbséga nok és a férfiak beszéde között, hogy a vonalas spektrum léceit a férfiaktnál 100 Hz, a noknél 200 Hz választja el egymástól (ezért pl. a nok magas hangon nem tudják az *u* hangot kiénekelni)
- tobábbi érdekesség, hogy a magánhangzók magasságát F_2 és nem pedig F_1 határozza meg[1]
- a magánhangzók hossza kb. 30-60 ms





[1] Tehát pl. az uhang F_1 frekvenciája nagyobb, mint ugyanazon hangmagasságban lévo i-é, de az F_2 már az i-nél magasabb.

Angol magánhangzók F_1, F_2 grafikonja



1.9.1.2. Nazálisok

- 250-300 Hz környékén van az F1, de magasabb frekvenciákon nincsenek formánsok
- sokkal kisebb az energiájuk, mint a magánhangzóknak

1.9.1.3. Likvidák

- van formáns struktúrájuk, de nem jellegzetes
- nagy szabadságfokú, hogy hol képzodik
- sokkal kisebb az energiájuk, mint a magánhangzóknak

1.9.2. Felpattanó zárhangok

- általános szerkezetük: néma fázis (60-120 ms), folytott zönge/zömgétlen (30 ms), zárfelpattanás (40 ms) hallható aspiráció/átvezetés a következo hangra
- a hang hosszítása a néma fázis hosszabbításával történik

1.9.3. Réshangok

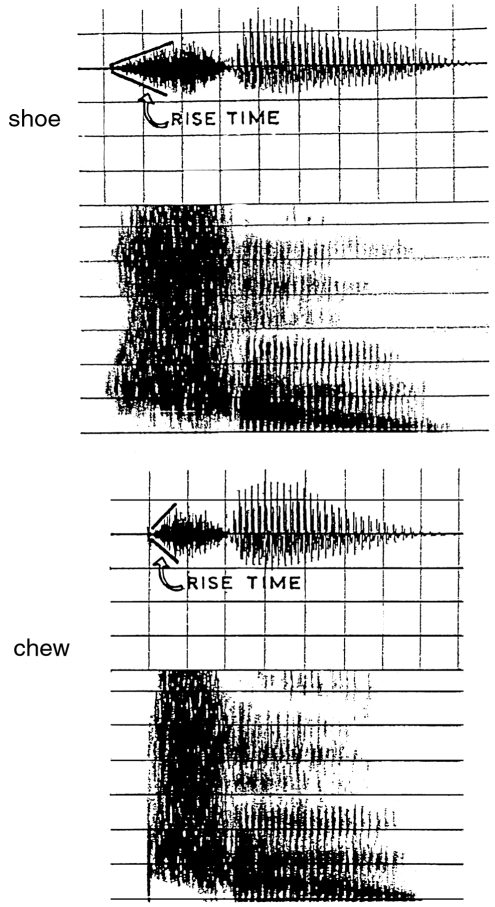
- lényeges, hogy frekvenciában 3,5 kHz felett van aspektrumuk (ezért pl. telefonban ezeket a hangokat nem tudjuk megkülönböztetni)
- ha a hang zöngés, akkor erre a spektrumra szuperponálódik rá egy formáns szerkezetu vonalas spektrum

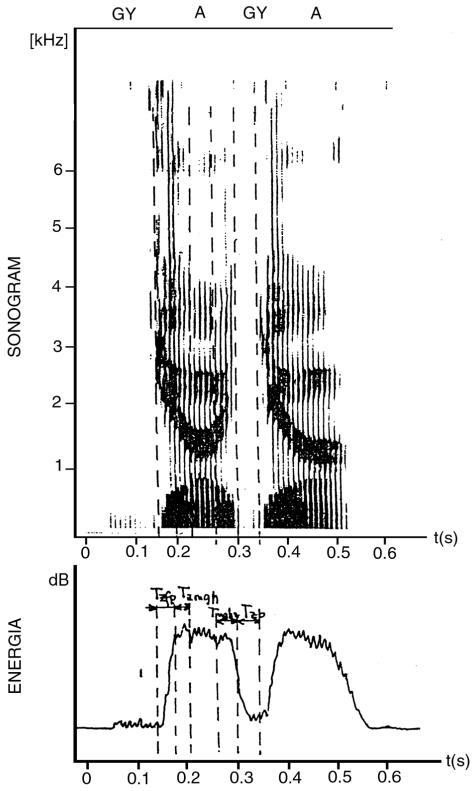
1.9.4. Zár-rés hangok (affrikáták)

- általános szerkezetük: néma fázis, folytott zönge/zöngételen, zárfelpattanás, réshang
- fontos az idoszerkeze: a néma fázis az zár-rés hangok elott rövidebb, mint a zárhangoknál és a réshang kialakulása sokkal gyorsabb (10-15 ms), mint a tiszta részhangé, továbbá a tiszta réshang tartási ideje sokkal kisebb (rövidebb)

1.9.5. Hangámenetek

- az ember a beszéloszerveit nem tudja ugrásszeruen változtatni, ezért beszéd közben átmeneteket képez a hangok között (végtelen sok félét)
- az átköztéseket alapvetoen a szomszédok határozzák meg, bizonyos esetekben az 1, 2 –vel után elhelyezkedo hangok is hatással vannak a kiejtett hangra
- tipikus vizsgált kombinációk: cv, vc, cvc, vcv[1]





1.9.5.1. Hosszan tartható hangok környezete

- (szomszéd) – (átmenet – tiszta fázis – átmenet) – szomszéd
  - átmenet: látszólag összevissza függvény, egyre jobban hsonlít a tiszta fázis jelalakjához
  - tiszt fázis: kvázi-periodikus jel
  - átmenet: többnyire amplitúdó csökkenés történik, de a legvégén itt is vanna torzulások, mint a kezdo átmenetnél

1.9.5.2. Locus

- 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 ot követo magánhangzó vagy ot megelozo mgánhangzó második formánsát a szóban forgó mássalhangzót jellemzo frekvenciára kényszeríti, ezek a locusok. Legjellemzobb locusok a felpattanó zárhangoknál vannak, ezek elég jól jellemezhetok a locusaikkal.

[1] c: consonant (mássalhangzó), v: vowel (magánhangzó)

1.10. A folyamatos beszéd akkusztikai szerkezete

1.10.1. Hangsúly

- a *hangsúly* az, hogy a mondanivaló függvényében ugyanazt a beszédhangot különbözoképpen ejtjük
- *intonáció*: $F_0(t)$, a zönge alapfrekvenciájánka változása az idoben
  - ezt szekundumos nagyságrendben érzékeljük[1]

1.10.2. Intenzitás

- rövid ideju energia (a jel által hordozott energia egy ún. kiablakolt energia)
- ugyanazt a magánhangzót egy kérdo mondatban más intenziással ejtjük, mint egy kijelento mondatabn

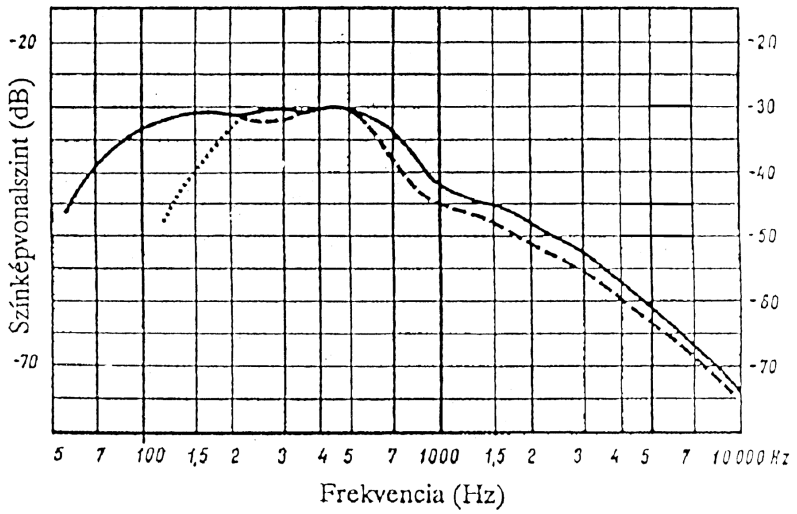
1.10.3. Ritmus

- a különbözo hangokat különbözo hosszúsággal ejtjük
- a magánhangzók tiszta fázisainak hosszával szabályozzuk

1.10.4. Statisztikai jellemzok

- stacionárius sztochasztikus folyamatként tekinve a beszédet
- tetszoleges helyen kiválasztott valószínuségi változó suruségfüggvényével tökéletesen jellemezheto egy stacionárius sztochasztikus folyamat, ez az ún. *amplitúdóeloszlás*
  - normáljuk ezt a suruségfüggvényt a saját szórására
  - a beszéd sokkal jobban feszíti az eszközök amplitúdótartományát
  - a beszédben kb. 30 dB dinamikakülönbség és még ennyi hangero
  - σ szórás a hangerore jellemzo
- ugyanezt a folyamatot a frekvenciatartományban vizsgálva a spektrálsuruség írja le jól
  - $10\lg\left(\frac{s(f)}{s(f_r)}\right)$, ahol s(f) a spektrálsuruség a kérdéses frekvencián, $s(f_r)$ pedig a spektrálsuruség a referenciafrekvencián

[1] Létezik mikrointonáció is (ez egy hangon belüli alapfrekvencia-változás), ettol lesz személyre jellemzo, természetes hangzás



62 magyar beszélô átlagos színképe. A folytonos vonal a férfiak és nôk burkológörbéjét jelzi. A nôi beszédszínkép ez alatt marad, pontozott vonal jelzi az eltérését. Ahol pedig a férfi beszédszínkép marad alatta, szaggatott vonal jelzi az eltérést.

2. BESZÉDKÓDOLÁS ÉS TÖMÖRÍTÉS

2.1. Bevezetés

- a beszéd folyamata analóg jel: idoben folytonos, értékkészlete folytonos
- a beszédfeldolgozásban digitális eszközökkel dolgozunk, tehát a beszédet bitek sorozatává kell alakítanunk, ennek módja: mintavételezés → kvantálás → kódolás[1]
- analóg jelsorozat bitsorozattá való átalakításakor az alábbi jellemzokre kell figyelni
  - huség: valamilyem matematikai mértékkel mérheto (pl. négyzetes hiba) vagy pszichoakusztai mérések eredménye alapján
  - sebesség
  - komplexitás: a megvalósíthatóság szempontjából érdekes
- ezek a paraméterek egymás ellen dolgoznak, u, a tömörítés ugyan a sebességre irányítja a figyelmet, de mindhárom egyaránt fontos.

2.2. Mintavételezés

Shannon-Kotyelnyikov: Ha egy jel B sávra korlátozott, azaz a jelben adott B frekvencia fölött nem fordul elo komponens, akkor $f_0 \geq 2B$ suruséggel vett mintáiból a jel egyértelmuen visszaállítható.

Visszaállításhoz $\frac{\sin x}{x}$ interpoláló függvényt alkalmazunk minden mintavételi pontban az adott minta értéke és a többi mintavételi pontban pedig 0. Probléma: ez egy matematikai függvény, nem megvalósítható.

2.2.1. PAM[2] típusú simító visszaállítás

- x(t) jelet $f_0 = 1/T_0$ frekvenciával mintavételezzük, és így eloáll $\{x_i\}$ jel. Legyen az elemi jelkelto m(t) súlyfüggvényu, kimenete pedig változzék az x_i-vel arányosan: $x_i m(t-iT_0)$. Vegyünk továbbá egy g(t) súlyfüggvényu szurot, amelyre ráeresztve az elemi jelkelto által generált jelet, egy $\tilde{x}(t)$ függvényt kapunk. Az átvitel akkor huséges, ha $x(t) = \tilde{x}(t)$.
- Legyen M(f) az m(t) és G(f) a g(t) Fourier transzformáltja.

Tétel: Egy B frekvenciasávra korlátozott jel T0 = 1/f0 idoközu mintáiból akkor állítható PAM típusú simító visszaállítással helyre, ha f0>2B és H(f) = M(f)G(f) -B és B között konstans[3], a mintavételi frekvenciák B sugarú környezetében 0, egyébként pedig tetszoleges.

a Shannon-tétel ennek egy speciális esete, amikor (B és f0 között kijelölve egy pontot: x) –x és x között H(f) értéke 1, azon kívül pedig 0. Az ilyen szögletes spektrum idotartománybeli megfeleloje:

$$h(t) = \frac{\sin x}{x} = m(t) * g(t) = \int_{-\infty}^{\infty} m(t)g(t-t)dt$$

- gyakorlati szempontból: megvalósíthatósági megfontolások – a gyakorlatban impulzusokat állítunk elo, melyeknek spektrum egy lefelé görbülo $\frac{\sin x}{x}$-hez hasonló jelalak. Ez az M(f). Ahhoz, hogy H(f) a tételnek megfelelo legyen, -B és B között G(f)-nek felfelé kell görbülnie, majd f_0-B-nél már közel 0[4] értéket kell képviselnie. (Általában f_0-B-ben a szuronek pólusa van)

[1] A kódolás nem feltétlenül egy mintát vesz alapul, lehetséges kódolás minták egy véges halmazára is.
[2] Pulse Amplitude Modulation
[3] Ha az elemi jelkelto elé egy T_0-as szorzót teszünk, akkor ez a konstans éppen 1 lesz.
[4] Frekvenciában nem tudunk tartósan zérus átvitelt biztosítani: 40 – 60 – 80 dB-nyit.

- további szempont, hogy a fáziskarakterisztika lineáris legyen, ami azt is jelenti, hogy a csoportfutási ido konstans. Meredek lefutások elott viszont a csoportfutási idonek mindig csúcsa van
- Manapság már inkább G(f)-et tervezik és ehhez alakítják m(t)-t.

2.2.2. Azonos mintájú jelek

- Vegyünk azonos mintájú, de különbözo idofüggvényu jeleket: $x_i(kT_0) = x_j(kT_0)$, $\forall k$. Van-e spektrális rokonság az azonos minták között?

Tétel: az azonos mintájú jelek halmozott spektrumai megegyeznek, azaz

$$\sum_{l=-\infty}^{\infty} x_i(f - lf_0) = \sum_{k=-\infty}^{\infty} x_j(f - kf_0)$$

ahol $\sum_{l=-\infty}^{\infty} x_i(f - lf_0)$ az $x_i(t)$ halmozott spektruma.

- A halmozott spektrum a képzés módja miatt periodikus. Az $-\frac{f_0}{2}, \frac{f_0}{2}$ közötti szakaszát nevezzük Nyquist-ekvivalensnek. A Nyquist-ekvivalens az azonos mintájú jelek közül a legkisebb sávszélességu. Ha a jel eredetileg teljesíti a mintavételi tétel frekvencia-feltételét, akkor a Nyquist ekvivalense önmaga.
- A minták egyértelmuen meghatározzák a halmozott spektrumot és a mintákra halmozott spektrumot határozzák meg egyértelmuen.
- A halmozott spektrum tulajdonképpen egy Fourier sor a frekvenciatartományon (innen is látszik, hogy ez periodikus jel):

$$\sum_{l=-\infty}^{\infty} x_i(f - lf_0) = T_0 \sum_{n=-\infty}^{\infty} x(nT_0) e^{-j2p\, f\, nT0}$$

2.2.3. PAM típusú simító visszaállító kimeneti jele

$$X(f) = T_0 \left[\sum_{n=-\infty}^{\infty} x(nT_0) e^{-j2p\, f\, nT0} \right] M(f)G(f) = \left[\sum_{l=-\infty}^{\infty} X(f - lf_0) \right] M(f)G(f)$$

és $\sum X(f - kf_0)$ azon jelnek halmozott spetruma, amelybol a minták származnak.

- A gyakorlati mintavételezés során x(t) halmozott spektruma H(f)-fel lép interakcióba.
- Helyes spektrumfeltétel esetén a halmozott spektrumban nincsenek átlapolódások, a mintavett jel spektruma megegyezik a Nyquist-ekvivalensével, a visszaállítás során így éppen az eredeti jelet kapjuk.
- Ha a spektrumfeltételt nem tartjuk be, akkor a mintavett spektrumok a halmozott spektrum képzése során összegzodnek, ebbol az összegzett spektrumból az eredeti nem állítható elo huségesen. Ilyen esetben törekedhetünk arra, hogy levágjuk azokat az széleket, ahol a kicsúcsosodások vannak, illetve a jelet elotte is sávkorlátozhatjuk.
- Idegen átlapolódás jelensége: ha a jelünk ugyan sávkorlátozott és betartja a frekvencia-feltételt, de a nagyfrekvenciás tartományokban van valami idegen jel, akkor ez a halmozás folyamán megjelenik a hasznos sávban és a jelbol többé nem távolítható el. Ezen hibák kiküszöbölésre bemeneti aluláteresztо szurot alkalmazunk, amely ugyan levágja a jel egy részét, viszont így a jel oly módon lesz sávkorlátozott, hogy betartja a frekvencia-feltételt, és ha a bemeneti szuro nem ideális voltából adódó torzításokat H(f)-fel kikompenzáljuk, akkor a nem levágott részeket torzításmentesen tudjuk átvinni. Tehát torzítás csak az átlapolódás megszuntetése miatti sávkorlátozásból adódik.

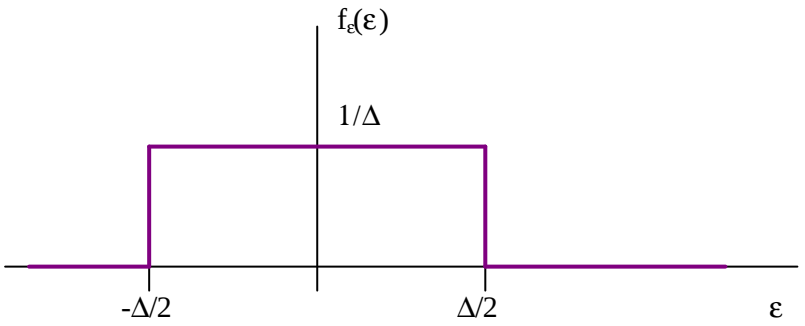
2.3. (Lineáris) kvantálás

a kvantálás általában a jelet nem idoben, hanem amplitúdó értékkészletben teszi diszkrétté

- többnyire mintavett jelre alkalmazzuk
- véges amplitúdó értékkészlet:
  - kvantálási szintek: a megengedett szintek
  - kvantálási lépcso: a megengedett szintek közötti lépcsök
- lineáris kvantálás esetén $\Delta_i = \Delta$ $\forall$i-re, és a kvantálási szabály az, hogy minden értéket a hozzá legközelebb eso kvantálási szinttel helyettesítjük.[1]

2.3.1. A kvantálás ára: a kvantálási zaj

Legyen e a kvantálási hiba értéke, $\hat{x}$ a kvantált és x a kvantálandó érték. Ekkor $\hat{x} = x + e$. Ésszeru modellt választva ε-ra, o egy valószínuségi változó $f_\varepsilon(\varepsilon)$ suruségfüggvénnyel.



A kvantálási hibát, mint számsorozatot tekintve: ε_i és ε_j idoben egymás után megjeleno valószínuségi változók. Ezek az ε értékek egymástól függetlenek, köztük korreláció nincs: $M(\varepsilon_i \varepsilon_j) = M(\varepsilon_i) \cdot M(\varepsilon_j)$ $\forall i \neq j$ esetén. Mivel az eloszlás szimmetrikus a 0-ra, ezért a várható érték 0.

A mintasorozat visszaállítása – $\hat{x} \rightarrow \tilde{x}(t)$ – során kizárólag lineáris muveleteket feltételezve: $x + e \rightarrow x(t) + e(t)$. Ez az ε(t) additív jelenség, zaj. A függetlenségbol következoen a számunkra fontos ($-f_0/2, f_0/2$) frekvenciasávban jó közelítéssel állandónak, tehát fehérzajnak tekintheto.

[1] A kvantálást 1938-ban Reeves „találta fel”. Észrevette, hogy a nem kvantált jelet átvíve a jelhez zaj adódik, és így a jel már nem állítható vissza. Ha azonban a zaj egy kvantált jelben okoz torzulást, és a hozzáadódó zaj kisebb, mint a kvantálási lépcso fele, akkor észreveheto, hogy hiba történt, és ennek megfeleloen újra kvantálhatunk. Ezt a folyamatot nevezik regenerálásnak. Manapság a kvantálás a bináris ábrázoláshoz és a kódolhatósághoz szükséges. A kvantálással való átalakítást az ADC (analóg/digitál konverter) végzi.

2.3.2. A kvantálási zaj teljesítménye

Mivel e egy sztochasztikus folyamatnak tekintheto, ezért

$$P_e = M(e^2) = \int_{-\infty}^{\infty} e^2 f_e de = \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} e^2 \frac{1}{\Delta} de = \left| \frac{1}{\Delta} \cdot \frac{e^3}{3} \right|_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} = \frac{\Delta^2}{12}$$

Ezzel azonban nem kaptuk meg a hasznos jel és a zaj viszonyát. Tekintsünk egy referencia jel/zaj viszonyt: vegyünk egy, a teljes kvantálási tartományt kitölto C amplitúdójú szinuszjelet. N szinten kvantálva N=2C. Ennek a szinuszjelnek a teljesítménye $P_s = \frac{C^2}{2}$. A kvantálás jel/zaj viszonya SNR (Signal to Noise Ratio).

$SNR = \frac{\frac{C^2}{2}}{\frac{\Delta^2}{12}} = 6\frac{C^2}{\Delta^2}$.Tehát N kvantálási szint esetén $\frac{C}{\Delta} = \frac{N}{2}$ *és* $SNR = \frac{3}{2}N^2 = \frac{3}{2}2^{2n}$, ha az N kvantálási szintet n biten kódoljuk. Ebbol már adódik, hogy $\mathbf{SNR^{[dB]}} = 10 \cdot \lg(\mathrm{SNR}) = \mathbf{1.74 + n \cdot 6.02\ dB}$

A kvantálás jel/zaj viszonya tehát kifejezheto a kvantálás során felhasznált kódszavak hosszával: 8 bites kódok esetén kb. 49.7 dB, 16 bites kódszavak esetén pedig kb. 97.74 dB.[1]

2.4. Logaritmikus (PCM) kvantálás

Megfigyelték, hogy az analóg telefonvonalon a távoli elofizeto hangja nagyon kis teljesítménnyel, míg a közeli elofizeto hangos (esetleg üvölto) hangja nagy teljesítménnyel szerepel. Ez a teljesítményviszony elérheti akár a 60-70 dB-t is. A tapasztalat azt mutatta, hogy n=12 biten kellene kvantálni egyenletes kvantálással ahhoz, hogy a távoli elofizeto hangja is hallható legyen.

Stevens azonban észrevette, hogy az emberi fül a nagy amplitúdók esetén kevésbé érzékeny a hibákra, és kimondta az Stevens törvényt, miszerint $\frac{dx}{x} = c$, azaz ha a hiba és amplitúdó aránya konstans, akkor az érzeti világunk egyensúlyban van. Tehát a kvantálási lépcsok a 0-tól elfele nonek. Az ilyen kvantálást *nem lineáris kvantálás*nak nevezzük.

Kvantálási karakterisztikának nevezzük azt a karakterisztikát, amely a nem lineárist lineárissá transzformálja (ez a lineáris kvantálás esetén egy egyenes).

$\Delta x/x = c_1$ és $\Delta y = C_2$. f(x) = ? Elosztva egymással a két egyenletet, majd átrendezve az oldakat, a következo összefüggést kapjuk:

$$f'(x) = \frac{c_3}{x},\ ahol\ c_3 = \frac{c_2}{c_1}\ és\ így\ f(x) = \ln(x) + c_4$$

Ezt tovább finomítva két szabvány alakult ki, az európai PCM karakterisztika (A-law) és az amerikai PCM karakterisztika (μ-law).

A 12 bites lineáris kvantálással 96 kbit/s, míg a logaritmikus kvantálással 64 kbit/s átviteli sebességre van szükség. A fenti tömörítési módszert hívják érzeti tömörítésnek.

[1] Az itt felsorolt értékekbol is látszik, hogy $\mathrm{SNR}^{[dB]} \approx 6n$ és $\mathrm{SNR} \approx 2^{2n}$.

2.5. Lineáris predikció

Motiváció: a beszéd 8kHz-es mintavételezése esetén sokszor fordul elo, hogy lét egymást követo minta nem nagyon különbözik egymástól. Tehát ha nem a mintát, hanem a minták különbségét kvantáljuk, akkor kevesebb biten tudjuk átvinni ugyanazt az információt (tömörítés). Legyen ξ_k a k-adik idopillanatban lévo minta. Ekkor

$\tilde{x}_k = \sum_{j=1}^{M} w_j x_{k-j}$, ahol w_j az ún. súlytényezo (weight factor)

Gyakorlatban ezeket a súlytényezoket és az így számított értékeket akkor használják, amikor a beszéd stacionáriusnak tekintheto (hasonló, és ezért ezek a súlytényezok néhány mintán keresztül érvényesek). A *w* súlytényezoket keretrol keretre a mintákból határozzuk meg. Egy keret 10-20 ms hosszú (ez 80-200 mintát jelent), pl. egy magánhangzó tiszta fázisa lehet egy keret.

A *w*-k eloállításánál arra a törekszünk, hogy az így kapott $\tilde{x}_k$ minél jobban megközelítse ξ_k-t. A predikció hibája ν_k, de mivel ezek valószínuségi változók, össze-vissza ugrálnak, ezért ezeket jól jellemezhetjük a négyzetes várható értékükkel. Tehát *w*-k akkor optimálisak, ha

$E = M(\tilde{x}_k - x_k)^2$ minimális.

$$E = M(x_k - \sum_{j=1}^{M} w_j x_{k-j})^2 = M\left(\sum_{j=0}^{M} w_j x_{k-j}\right)^2 .$$

$$E = M\left(\sum_{j=0}^{M} (w_j x_{k-j}) \cdot (w_j x_{k-j})\right) = M\left(\sum_{i=0}^{M}\sum_{j=0}^{M} w_i \cdot x_{k-i} x_{k-j} \cdot w_j\right) = \sum_{i=0}^{M}\sum_{j=0}^{M} w_i \cdot M(x_{k-i} x_{k-j}) \cdot w_j$$

A szorzatok eredményét össze kell adni és el kell osztani. De ha a folyama ergodikus, akkor 1 folyamat is magán hordozza a sokaság tulajdonságait, tehát a k-val is végigfuttatható. Így az összeg az i-j távolságra lévo minták átlagával közelítheto.

Véve tehát $R_{ij} = R_{ji} = M(x_{k-i} \cdot x_{k-i})$ *autokorrelációs függvény*t, a fenti összefüggés így alakítható át:

$$\sum_{\substack{i=0 \\ i\neq m}}^{M}\sum_{\substack{j=0 \\ j\neq m}}^{M} w_i R_{ij} w_j + w_m \sum_{\substack{j=0 \\ j\neq m}}^{M} R_{ij} w_j + \left(\sum_{\substack{i=0 \\ i\neq m}}^{M} w_i R_{im}\right) w_m + w_m R_{mm} W_m$$

$$\frac{\partial E}{\partial w_m} = \sum_{\substack{j=0 \\ j\neq m}}^{M} R_{mj} w_j + \sum_{\substack{i=0 \\ i\neq m}}^{M} w_i R_{im} + 2 w_m R_{mm}$$

$$\frac{\partial E}{\partial w_m} = 2\sum_{j=0}^{M} w_j R_{mj} = 0$$

$$\sum_{j=1}^{M} w_j R_{mj} = R_{m0} = R_{0m}$$

A harmadik sorában a deriváltat azért tettük 0-vá ,mivel a minimalizáláskor a j-nek nem volt igazi változótartalma, azt mi definiáltuk –1-nek. Végül valójában M db. egyenletünk van. Ezeket felírva w_1, w_2 … w_M-re:

$$\begin{bmatrix} R_{11}w_1 + R_{12}w_2 + \ldots + R_{01} \\ \ldots \\ R_{M1}w_1 + R_{M2}w_2 + \ldots + R_{0M} \end{bmatrix}$$

M egyenlet és M ismeretlenes egyenletrendszer. Ez a lineáris predikció alapegyenlete. R-et a korrelációs mátrixnak nevezzük. A korrelációs mátrix további tulajdonságai:

- R_{ii} – k függetlenek i-tol, tehát $R_{11} = R_{22} = \ldots = R_{MM}$;s ez legyen R_∞.
- Ez a tujakdons'g tov'bbra is igaz a foátlóval párhuzamos átlókra: $R_{i-j} = R_{k-l}$, ha i-j = k-l. Ez az ún Toepliz mátrix.

$$\underline{\underline{R}} = \begin{bmatrix} R_{11} & R_{12} & \ldots & R_{1M} \\ R_{21} & R_{22} & \ldots & R_{2M} \\ \ldots & & & \ldots \\ R_{M1} & R_{M2} & \ldots & R_{MM} \end{bmatrix} = \begin{bmatrix} R_\infty & R_1 & & R_{M-1} \\ R_1 & R_\infty & & \\ & & & R_1 \\ R_{M-1} & & R_1 & R_\infty \end{bmatrix}$$

$$\ldots\ldots \text{és} \ldots \underline{b} = \begin{bmatrix} R_{01} \\ R_{02} \\ \ldots \\ R_{0M} \end{bmatrix}$$

Vagyis a lineáris predikció alapegyenlete tömör formában a következoképpen írható fel:

$$\underline{\underline{R}} \cdot \underline{w} = \underline{b} \cdots$$

$$\text{és ennek } \left|\underline{\underline{R}}\right| \neq 0 \text{ esetén } \exists \text{ megoldása } \underline{w} - \text{re,}$$

$$\text{hogy } E = M(x_k - \tilde{x}_k)^2 \quad \min \text{imális.}$$

Az egyenletrendszer megoldása nem egyszeru, de léteznek rá algoritmusok:

- Durbin
- Levinson
- rekurzív algoritmus
- adaptív algoritmusok

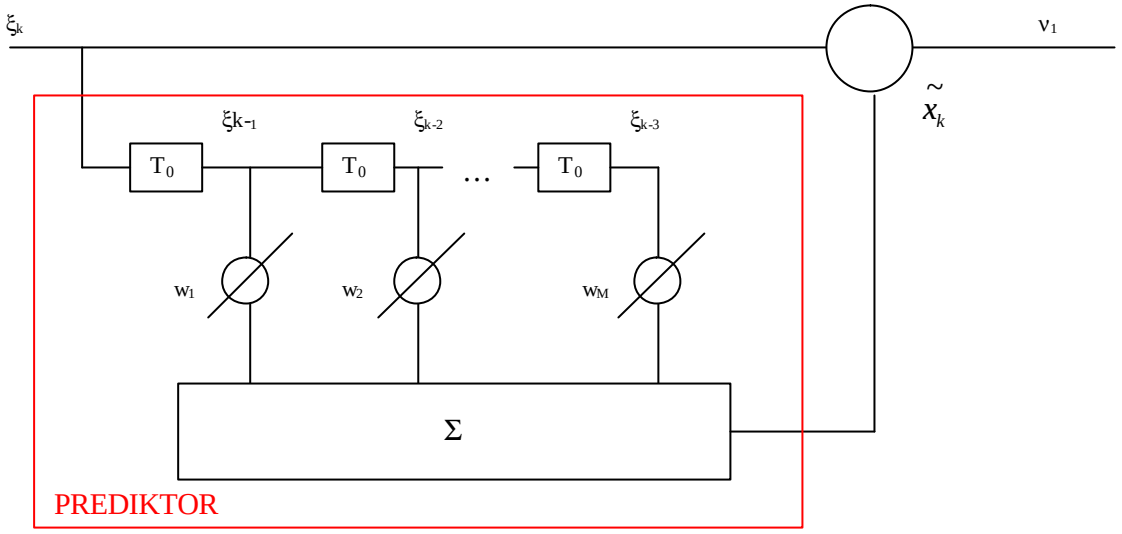
$$n_k = x_k - \tilde{x}_k$$

Ha a közelítés jó, akkor n_k kicsi és ekkor 1-1 n-beli kódolási minta kvantálásához kevesebb kvantálási szint szükséges.

- x kvantálása: n2 bit
- n kvantálása: n1 bit
- n1<n2

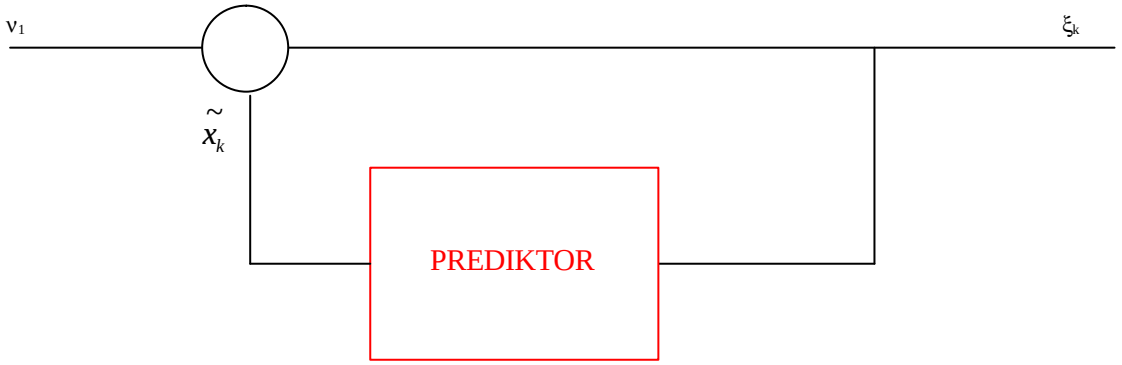
Kérdés: ez a tömörítés alkalmas-e arra, hogy az eredeti jel pontosan vagy elfogadható hibá van visszaállítható legyen?

2.5.1. Tömörítés és visszaállítás



- a tömöríto: transzverzális (digitális) szuro
- impulzusválasz függvénye: véges, azaz FIR (Finite Impulse Response)
- ez a lineáris predikció beszéd ananlízis modellje

A visszaállítás:

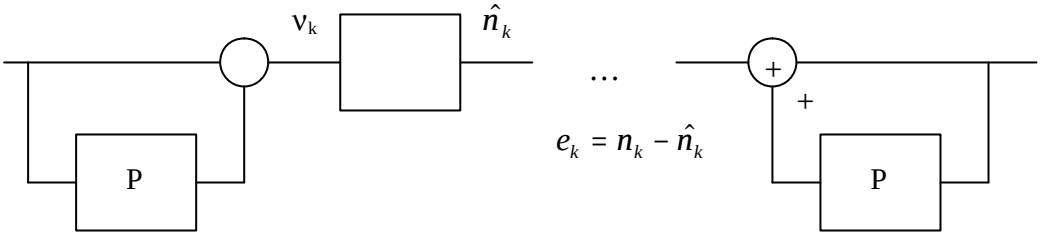


- A lineáris predkitív beszékkódolás szintézi modellje
- A bemenet rögtön megjelenik a kimeneten + visszacsatolás
- 1 bemeno impulzusnak elvileg végtelen válasza van
- IIR (Infinite Impulse Response)

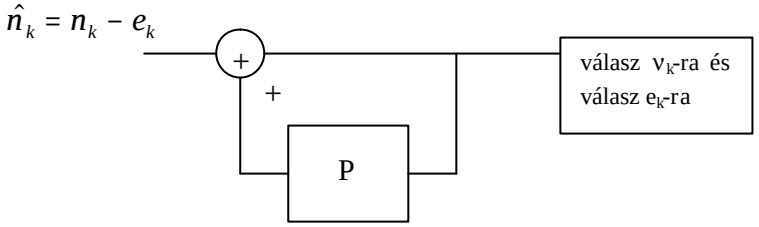
Fizikai értelem:

- másik, jól szegmentált magánhangzó esetén hasonló ν-ket kapunk, de a w_i-k nagyon különböznek, jellegre azonban hasonlóak
- az egészet helyettesíthetjük azzal, mintha a rendszert a zönge ütemében gerjesztenénk
- a szintézis modell ehhez a szemlélethez hasonlóan muködik
- más, pl. hosszan tartható zöngétlen hangok esetében (f, s) ez a gerjeszto hang fehérzajszeru, véletlenszeru, hasonló, mint a vokális traktusban (itt is véletlenzaj-generátor, turbulencia, leszakadó levegorészecskék, stb.)

2.5.2. A kódoló gyakorlati megvalósítása

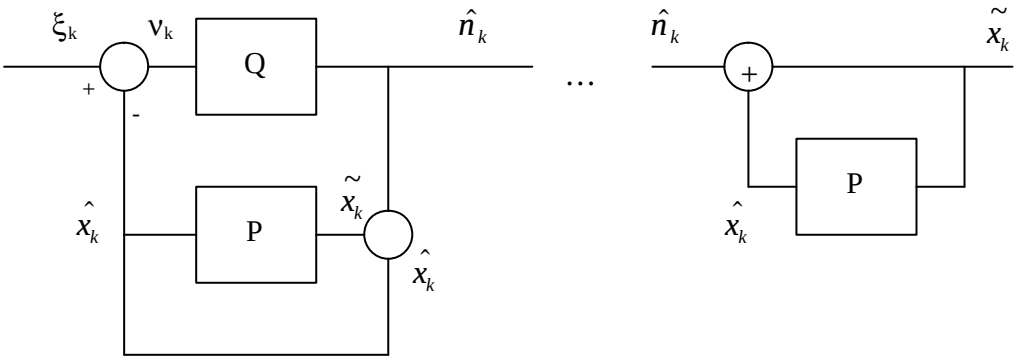


Mi lesz a kvantálási hibával a visszaállítás során?



- e_k: a tényleges és a kvantált érték közötti különbség, a kvantálási hiba
- ha csak egy ilyen is van, akkor a vevoben lévo IIF szuro miatt végtelen választ ad
- állandó hibát okoz, halmozódik a kvantálási hiba, egy ido után kitér a dekódoló tartományból
- minden hibát, amit ν_k elszenved, a kimeneten halmozottan kapunk vissza

A kvantálási hibát azonban kézben lehet tartani: valójában a kódolóba is beépítjük a dekódolót.



$$x_k - \tilde{x}_k = \hat{x}_k + n_k - \left(\hat{x}_k + \hat{n}_k\right) = n_k - \hat{n}_k$$

- vagyis a megmaradó hiba csak a tényleges kvantálási zaj
- a bemenet és a kimenet csak a kvanálási hibában különbözik
- a k-adik idopillanatban lévo hiba *kizárólag* a k-adik idopillanatban elkövetett hibától függ (emlékezetnélküli csatorna)

2.5.3. Lineáris predikció a gyakorlatban

- a jelet 10-15 ms-os darabokra bontjuk (ezek az ún. keretek)
- meghatározzuk ν-t
- meghatározzuk azt a 10 LPC együtthatót a vevonek (Ezek a w-k)
- a túloldalon visszaállított jelbol valamilyen trükkel megpróbáljuk meghatározni a predikciós együtthatókat (vagy eltároljuk, ha arról van szó)

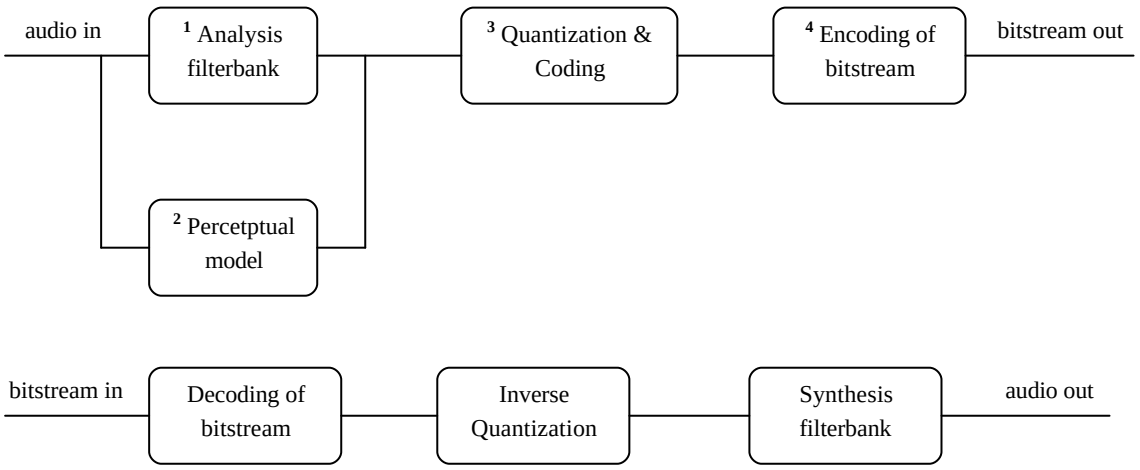
Trükkös esetek

- ν-k közül a két legnagyobbat választjuk ki vagy azt a hármat, amely egymás mellett a legnagyobb
- elképzelünk 1024 hibasorozatot, a tényleges ν-k helyett ebbol a készletbol visszük át azt, amelyik valamilyen értelemben a legjobban hasonlít a tényleges hibasorozatra (CELP)

2.6. Érzeti (részsávos kódolás) – percetptual (subband) coding

2.6.1. Frekvenciamaszkolási jelenség

- minden sávra megállapítjuk, hogy milyen energiájú összetevok vannak a jelben
- az elfedett összetevok kihagyása
- a kvantálási zaj növelésének lehetosége: úgy kvantálunk, hogy a kvantálási zaj ne legyen nagyobb, mint az elfedési szint (itt a tömörítési lehetoség)



1. A bemenetre kerülo jelet összetevokre bontja; elony, hogy a kisebb frekvenciájú jelet kisebb frekvencián kell mintavételezni, lejátszani (szurosor)
2. Érzeti modell: elfedési görbe meghatározása
3. Kvantáló és kódoló: az összetevok kvantálása több lépcsoben
4. bitfolyam kódolása szabványos formátumra (mintavételi frekvencia, szurokomponensek)

Elony az LPC-vel szemben:

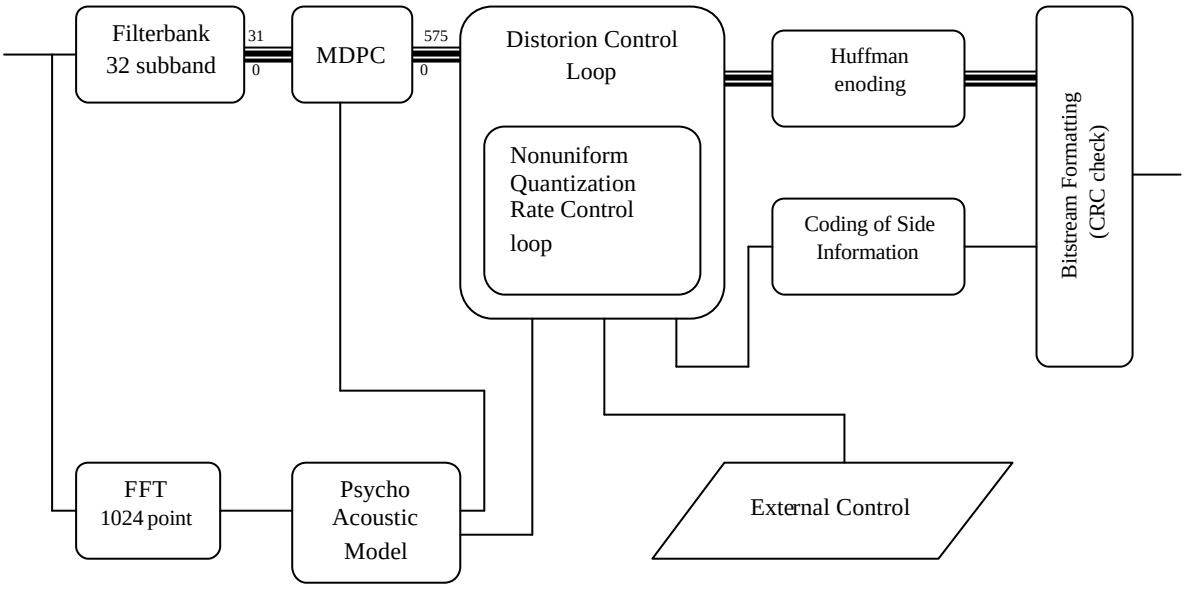
- az LPC-ben át kell vinni a súlyokat, a hibát és a szegmentálást, és egy szegmensben stacionáriusnak tekintjük a beszédet
- az új módszerrel nemcsak beszédet, hanem más audio jelet is át lehet vinni

2.6.2. Motion Pictures Expert Group

- hivatalos neve: ISO/IEC JTC/SC2 9/WG11
- Feladatok:
  - digital audio bradcasting (DAB)
  - ISDN
  - tárolás
  - DVB, HDTV

- Internet streaming
- hordozható audio mp3-lejátszó
- audio filecsere

- a fentiekbol csak a hangkódolással foglalkozunk
- különbözo minoségi szintek ugyanazon elv alapján (layernek hívják oket)
  - MPEG1 – 1992 (192 … 32 kb/s)
  - MPEG2 – 1994 (újabb frekvenciák: 16, 22,05, 24 kHz)
  - MPEG3 – HDTV-hez készült volna, de visszatértek az MPEG2-höz
  - MPEG4 – 1998
    újdonság: nem a tömörítés hatékonyságát javítják, hanem új szolgáltatásokat hoznak be (interaktív TV)
    hangkódolási tartomány: 2kb/s ----
  - MPEG7 – 2001? (kidolgozás alatt, tartalom-reprezentációs szabvány)



- bemenet: Digital Audion Signal (PCM 768 kb/s)
- kimenet: Coded Signal (2…192 kb/s)
- Filterbank: nagyjából a kritikus sávok szerinte felbontja a jelet, minden sávot további 18 részsávra lehet bontani, így lesz 0…575 sáv
- MPC: Modified Discrete Cosine Transform
- segédinformáció: pl. milyen Hufman-táblát használjunk

A szabvány a dekódert írja le

- nyílttá válik a lehetoség a kódoló implemetálására (ugyanazon dekódokó jobb kódolóval jobb minoséget produkál)
- változó bitsebesség (lehetséges az is, hogy bizonyos részeket más bitsebességgel viszünk át)
- többféle forrást kell tudnia: mono, sztereo hagyományos és kombinált, kétcsatornás

Problémák

- visszaállított jel: csendes szakaszban is megjelenik valamekkora jel, ez a zaj a pre-echo jelenség (nem kauzális a kódoló)
- ha úgy tudunk szegmentálni, hogy az idoelfedési jelenség miatt nem halljuk, azzal csökkenthetjük a hibát

3. BESZÉDVÁLASZÚ RENDSZEREK

3.1. Gépi beszédkeltés alapfogalmai: három kategóriát különböztetünk meg

3.1.1. Kötött szókészlet

- tudjuk, hogy a rendszernek mit kell majd mondania
- állandó üzenet („a hívott szám nem elérheto”, kiterjesztett magnetofon)
- változó elemek
  - primitív: „Önnek üzenete érkezett 2000 május”
  - bonyolultabb: „A hívott szám megváltozott, az új szám: 325-29-48”

3.1.2. Kötetlen szókészlet (text to speech, szövegfeldolgozó)

- gyakorlatilag ilyen nincs
- széles szókinccsel kell rendelkeznie kiinduló állapotban, és tetszolegesen bovítheto
- ha tudjuk a tematikát, kifejezéseket, akkor meg lehet tanítani

3.1.3. Vegyes rendszerek

- vannak állandó üzenetek (ezeket nem célszeru TTS-sel megoldani, mert fárasztó)
- vannak változó üzenetek

szókészlet * minoség = konstans

3.2. Kötött szókészletu rendszerek tervezési szempontjai

3.2.1. Tematika felderítése

- az adott rendszerven mik azok az információk, melyeket el kell juttatni a felhasználóhoz
- mik ezeknek a módjai
- a felhasználók figyelembevétele (kezdo + profi különbözo)

3.2.2. bemondandó szöveg tervezése

3.2.3. szótárkészlet kialakítása

- az elozovel szinkronban
- kompromisszum a minoség és a bonyolultság között
- szótárelemek számára algoritmus

3.2.4. bemondó választása

- akusztikai arculat

3.2.5. akusztikai adatbázis elkészítése

- felvétel készítése
- elemek kivágása és feldolgozása

3.2.6. rendszerbeillesztés

3.3. Konkrét példa:

- többnyelvu számbemondó tervezése és megvalósítása
- Hagyományos megoldás: írásnak megfelelo számelemek összefuzése szünetekkel
- Pl: 125000 :

(English)	one	hundred	and	twenty	five	thousand
(German)	ein	hundert	fünf	und	zwanzig	tausend
(Hungarian)	száz	huszon	öt	ezer		
(Portuguese)	cento e	vinte	cinco	mil		

- Magyar: 25 db elem, portugál: 53 db elem

	Basic element	English	German	Hungarian	Portuguese
1.	1	one [ω℘ν]	ein [αι ν]	egy [ε϶]	um [ũ]
2.	1--	--	eins [αι ν σ]	--	--
3.	1--	--	eine [αι ν ↔]	--	--
4.	2	two [τυ:]	zwei [τσ ω αι]	kettõ [κΕτ:Ο:]	dois [dojʃ]
5.	3	three [Τρι:]	drei [δ ρ αι]	három [ηα:ρομ]	três [treʃ]
6.	4	four [ϕ :]	vier [ϕ ι:]	négy [νε:϶]	quatro [kwatru]
7.	5	five [ϕαιϖ]	fünf [ϕ Φ ν ϕ]	öt [Οτ]	cinco [s ĩ ku]
8.	6	six [σικσ]	sechs [ζ Ε κ σ]	hat [η τ]	seis [sEjʃ]
9.	7	seven [sevn]	sieben [ζ ι: β ν]	hét [ηε:τ]	sete [σΕτ↔]
10.	8	eight [eit]	acht [α ξ τ]	nyolc [⁀ολ τσ]	oito [ojtu]
11.	9	nine [ναιν]	neun [ν φν]	kilenc [κιλεν τσ]	nove [ν ⌠ϖ↔]
12.	10	ten [ten]	zehn [τσ ε: ν]	tíz [τι:ζ]	dez [δΕʃ]
13.	10x			tizen... [tizen]	
14.	11	eleven [ιλεϖν]	elf [Ε λ ϕ]	--	onze [⌡ ⌡↔]
15.	12	twelve [τω↔λϖ]	zwölf [τσ ϖ ↵ λ ϕ]	--	doze [δ⌡ ⌡↔]
16.	13	thirteen [Τ↔:τι:ν]	dreizehn	--	trese [τρε⌡↔]

17.	14	fourteen	vierzehn [ϕ ι ρ τσ ε: ν]	--	catorze [κατορ⏌↔]
18.	15	fifteen	fünfzehn	--	quinze [κ i͡ ⏌↔]
19.	16	sixteen	sechszehn	--	dezasseis [δ↔⏌ασΕφ∫]
20.	17	seventeen	siebzehn	--	dezassete [δ↔⏌ασΕτ↔]
21.	18	eighteen	achtzehn	--	dezoito [δ↔⏌οφτυ]
22.	19	nineteen	neunzehn	--	dezanove [δ↔⏌αν ⌠ϖ↔]
23.	20	twenty [twenti]	zwanzig [τσ ϖαν τσ ι X]	húsz [ηυ:σ]	vinte [ϖ i͡ τ↔]
	Basic element	English	German	Hungarian	Portuguese
24.	2x			huszon.. [ηυσον]	vinte e [ϖ i͡ τφ]
25.	30	thirty	dreizig	harminc [η ρμιν τσ]	trinta [τρ i͡ τα]
26.	3x				trinta e [τρ i͡ ταφ]
27.	40	forty	vierzig	negyven [νεϑϖεν]	quarenta [κωαρ ẽ τα]
28.	4x				quarenta e
29.	50	fifty	fünfzig	ötven	cinquenta [σ i͡ κω ẽ τα]
30.	5x				cinquenta e
31.	60	sixty	sechzig	hatvan [η τϖ ν]	sessenta [σ↔σ ẽ τα]
32.	6x				sessenta e
33.	70	seventy	siebzig	hetven	setenta [σ↔τ ẽ τα]
34.	7x				setenta e
35.	80	eighty	achtzig	nyolcvan	oitenta [οφτ ẽ τα]
36.	8x				oitenta e
37.	90	ninety	neunzig	kilencven	noventa [νυϖ ẽ τα]
38.	9x				noventa e
39.	100	hundred [η ℘νδρ↔ δ]	hundert [ηϒνδ τ]	száz [σα:ζ]	cem [σ ẽ i͡]
40.	1xx				cento e [σ ẽ τυφ]
41.	200				duzentos [δυ⏌ ẽ τυ∫]
42.	300				trezentos [τρε⏌ ẽ τυ∫]

43.	400				quatrocentos [κωατρυσσ ẽ τυ∫]
44.	500				quinhentos [κι ⌡ẽ τυ∫]
45.	600				seiscentos [σΕφ∫σ ẽ τυ∫]
46.	700				setecentos [σΕτ↔σ ẽ τυ∫]
47.	800				oitocentos [οιτυσ ẽ τυ∫]
48.	900				novecentos [ν ⌠ϖ↔σ ẽ τυ∫]
49.	1000	thousand [Ταυζνδ]	tausend [τ αυ ζ ν τ]	ezer [ΕζΕρ]	mil [mil]
	Basic element	English	German	Hungarian	Portuguese
50.	1000x				mil e [milj]
51.	1000000	million [μιλφ↔ν]	million [μ ι λ φ ο: ν]	millió [μιλιο:]	milhão [μιλ×© ũ]
52.					milhão e [μιλ×© ũ φ]
53.					milhões [μιλ×ο ĩ ∫]
54.					milhões e [μιλ×ο ĩ ⌡φ]
55.		billion [βιλφ↔ν]	milliarde [μιλφαρδ↔]	milliárd [μιλια:ρδ]	bilião [βιλι© ũ]
55.					biliões [βιλι ĩ ∫]
56.	0	O [ou]	--	--	--
57.	0-	zero [ζι↔ρου]	null [νυλ:]	nulla [νυλ:]	zero30 [ζερυ]
58.		and [Θνδ]	und [υντ]	--	e [φ]

Természetes kiejtéshez biztosítani kell:

- folyamatos kiejtés, helyes pozíciójú és hosszúságú szünetekkel
- a számelemek kiejtési helytol függo idoszerkezete
- spektrális és intenzitás folytonosság (koartikuláció figyelembe vétele) az elemhatárokon
- szóhangsúlyok és alapfrekvencia változások helyessége

3.3.1. Folyamatos kiejtés

A megfelelo helyeken, megfeleló hosszúsággal beiktatott szünetekkel, a 2, 3, 4, szempontok szerint kiválasztott elemek folytonos összefuzése (vágás nullátmenetnél negatívból pozitívba)

3.3.2. A számelemek kiejtési helytol függo idoszerkezete

Kezdo (B, beginning, pl. 1234567), középso (M, middle, 1231567), záró (L, last, 1234561) elem szükséges a többi szempont szerint kiválasztott minden elembol (elvileg).nagyszámú (közel ezer) kimondott szám vizsgálata alapján

3.3.3. Spektrális és intenzitás folytonosság (koartikuláció figyelembe vétele) az elemhatárokon

Minden elemre hat az elozo és a következo elem

Lehetséges pozíciók:

- Egyedül áll (6)
- Felsorolás (12, 2 56.)
- Elso (elemXXX)
- Belso (XXXelemXXX)
- Záró (XXXelem)

3.3.3.1. Az 1 példája

3.3.3.1.1. Angolul *one.*

*one*XXX után *hundred, thousand, million, billion,* (pl., 1100)
XXX*one* elott *thousand, million, billion, and, twenty, thirty.... ninety* (pl., 1100, 101, 21).
one elemkészlet:
(1) szabály: *one* felsorolásban vagy egyedül 1, 2, 3
*one*XXX esetekben
(2) szabály: *one* (*one hundred*), a (h) módosítja az (n)-et,
(3) szabály: *one* (*one thousand*) az (n) és a (t) azonos artikulációs bázisú, ezért az (n) rövidebb lesz,
(4) szabály: *one* (*one million*), az (n) (m)-be megy át,
XXX*one* esetekben
(5) szabály *one* (*thousand one, hundred and one*, etc.) zárhang és (v) találkozása,
(6) szabály: *one* (*twenty one,* etc.) i és (v) találkozása ,
XXX*one*XXX
(5) + (2), (5) + (3), (5) + (4), (6) + (3), (6) + (4)

Összesen: 11 (1+3+2 +5) elméleti lehetoség.
A hagyományos módszer minden elemére elvégezve a fenti elemzést, a spektrális és intenzitás folytonosság biztosítható.

3.3.3.1.2. Németül:

(1) szabály: *Ein* ha a szám 1-nél nagyobb, *eins* ha egyedül áll (pl., 1, 2, 3), *eine* pl. *eine million* és *eine DM*.
(2), (3), és (4) változatlan, mert az *ein*XXX és a *one*XXX kategóriái megegyeznek
XXX*ein* különbözik
(5) szabály: *hundert ein, tausend ein, milliard ein*, zárhang és magánhangzó találkozása,
(6) szabály: *million ein,* nazális és magánhangzó találkozása.
XXX*ein*XXX
(5) + (2), (5) + (3), (5) + (4), (6) + (2), (6) + (3)
Összesen: 13 (3+3+2 +5) elméleti lehetoség.

3.3.3.1.3. Magyarul *egy*:

(1) szabály: *egy* egyedül áll (1, 2, 3 stb.),
*egy*XXX
(2) szabály: *egy millió* és *egy milliárd*)
(3) szabály: *egy ezer*, pl. 31000), zöngés alveolo-palatális zárhang és magánhangzó találkozása,
(4) szabály: *egy száz* pl. 3125000, zöngés alveolo-palatális zárhangot zöngétleníti a *száz sz* hangja,
XXX*egy*
(5) szabály: ....*n egy*, pl. 51, 61, 71, etc.) a nazális hang módosítja az *e*-t,
(6) szabály: *millió egy*, pl. 5000001) magánhangzó-magánhangzó kapcsolat.
XXX*egy*XXX
(5) + (2), (5) + (3), (6) + (3), (6) + (4)
Összesen: 10 (1+3+2 +4) elméleti lehetoség.

3.3.3.2. A legfontosabb regresszív koartikulációs szabályok

az elozo elem utolsó hangja	az alábbira változik, ha	a következo elem elso hangja
b, d, g, v, z, 3	p, t, k, f, s, S	zöngétlen
ts	ts felpattanás (burst) nélkül	s
t	t felpattanás (burst) nélkül	n
n	n(k)	k
n	n(h)	h
n	nn	n
n	ɲ:	ɲ
n	m	m, b, p
magánhangzó	átmeneti szakasz	magánhangzó
magánhangzó	palatalizált átmeneti szakasz	palatális

A legfontosabb progresszív koartikulációs szabályok

ha az elozo elem utolsó hangja	és a következo elem elso hangja	akkor a következo elem leso hangja az alábbira változik
nazális	magánhangzó	nazalizált átmeneti szakasz
palatális	magánhangzó	palatalizált átmeneti szakasz
magánhangzó	magánhangzó	átmeneti szakasz

previous element in concatenation	element of the inventory	next concatenated element	example number
--	1(εϑ)	--	1
--	1(εϑ)(m)	millió, milliárd	1564322
millió	1(o)(εϑ)	--	3000001
1000 (EζEρ)	1(εX)	100 (σα:ζ)	1100
..(an), ..(en), ..(on)	1(n)(eJ)	1000 (J)(EζEρ)	51000
..(an), ..(en), ..(on)	1(n)(eJ)	millió, milliárd	61000000
any element	2(κEτ:O:)	100 (σα:ζ), millió	200, 312
any element	2(κEτ:O:)	1000 (O:)(EζEρ)	2000
any element	3(ηα:ρομ)	100, 1000, millió	300, 3000
any element	4(νε:ϑ)	1000 (J)(EζEρ)	4555
any element	4(νε:ϑ)	millió (J)(μιλιο:)	4000000
any element	4(νε:X)	100 (σα:ζ)	400
--, 30, 100, 1000	5(Oτ),	100, 1000, millió	535, 5000
..(ν), ..(en), ..(on)	5(n)(Oτ)	100, 1000, millió	65, 75, 25
any element	6 (η τ),7(ηε:τ)	100, 1000, millió	600, 700
1000, millió	8( )oλ τσ) without burst in [τσ]	100	812
10--90, 100	8( )oλ τσ)	1000, millió	8000, 8000000
1000, millió	9(κιλEν τσ) without burst in [ts]	100	900
10-90, millió	9(κιλEν τσ)	1000, millió	59000
any element	10(τι:ζ)	(EzEr),millió,	510000
--, 100, 1000, millió	variants for ending -Eν in numbers1x, 4x, 5x, 7x, 9x	1, 5,	11, 115
" "	-Eν)(k)	2, 9	12, 142, 79
" "	-Eν)(h)	3, 6, 7	13, 53
" "	-Eν)(n)	4	14, 94
" "	-Eν)( )	8	18, 98
" "	20 (ηυ:σ)	(EζEρ), millió,	20000
" "	variants for ending -on in numbers 2x, (-on)	1, 5	21, 125
" "	-on)(k)	2, 9	22, 122
" "	-on)(h)	3, 6, 7	23, 1223
" "	-on)(n)	4	24, 224
" "	-on)(ø)	8	28
" "	30 (η ρμιν τσ)	1,2,3,4,5,6,7,8,9, (EζEρ), millió	30256
" "	variants for ending - ν in numbers 6x, 8x, - ν)	1, 5	61, 185
" "	- ν)(k)	2, 9	62, 289
" "	- ν)(h)	3, 6, 7	63, 187, 666
" "	- ν)(n)	4	64, 164
" "	- ν)( )	8	168, 968
1,2,3,4,5,6,7,8,9, 1000	100 (σα:ζ)	1,4,5,8,40,50,80, 1000,millió	1000010í

" "	100 (σα:σ)	2,3,6,7,9,10,1x,20, 2x,,30,60,70,90	102
5,6,7,8,9,100	1000 (ΕζΕρ)	any number element	5001
2	(O:)(ΕζΕρ) 1000	" "	2000
1, 4	(ϑ)(ΕζΕρ)1000	" "	4000
40,50,60,70, 80,90	(n)(ΕζΕρ) 1000	" "	50000
3	(m)(ΕζΕρ)	" "	3000

3.3.3.2.1. Szóhangsúlyok és alapfrekvencia változások helyessége

sample number	pronounced style	comment
121	o n e h u n d r e d a n d t w e n t y o n e. AB N N AM AL	.=full stop AL= accent and falling intonation in the last item
2151	t w o t h o u s a n d o n e h u n d r e d a n d f i f t y o n e. AB N AM N N AM AL	AB, AM= accents in the number

A számok kimondásakor több hangsúly is megjelenik.

- AB: kezdo hangsúly
- AM: közbenso hangsúly
- AL: záró hangsúly, eso intonáció
- N: semleges, hangsúlytalan elemek

Szerencsére a helyes idotartamot biztosító elemek (B, M, L) megfelelo tervezés esetén magukban hordozzák a helyes hangsúlyt is.

- Ha a számelem a mondat végén áll, (pl. Az ön számlájának egyenlege: **53424** forint) eso jellegu intonációja lesz.
- Ha a mondat közepén helyezkedik el, (pl. Az ön számláján **53424 forint** összegu tranzakció valósul meg.) a számelem intonációja laposabb, lebegobb).

3.3.4. A számkimondó megvalósítása

Elozmény:

- az elemi (hagyományos) építokockák, számelemek meghatározása
- a kimondási szabályrendszerek (idotartam, koartikulációs, hangsúly és intonáció) meghatározása

3.3.4.1. A felolvasandó szöveglista meghatározása

- Vivoszöveg kialakítása az építokockák és a szabályrendszer alapján
- Example of determining the list of number elements and the source from where they will be cut out (for English)

number element	position	rule type	example of the recorded sample number which contains the element
one	B, L	(1)	1, 2, 1. (with pauses)
one(h)	B, M	(2)	121, 2151
one(t)	B, M	(3)	1121, 2001121
one(m)	B, M	(4)	1121151
(d)one(h)	M	(5), (2)	1122
(d)one	L	(5)	121101.
(ty)one(t)	M	(6), (3)	531231
(ty)one	L	(6)	541.

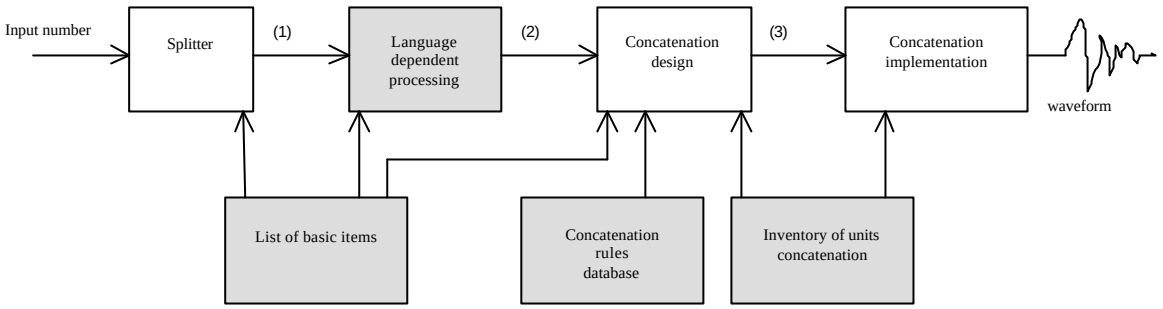
3.3.4.2. A felolvasandó szöveg felvétele

- Minden elemet a megfelelo vivoszövegben kell felolvasni. A vivoszöveget célszeru redundánsra tervezni (minden elem legalább kétszer forduljon elo).
- Az egyes elemek között kb. 2 sec szünetet célszeru tartani.
- Nagyobb egységenként (pl. oldalanként) érdemes hosszabb szünetet tartani.
- Az oldal megkezdése elott az elozo oldal végének meghallgatása.
- Összpontosítás az egyenletes hangmagasssság, hangero és beszédsebesség biztosításához.

3.3.4.3. A hangelemek kivágása a felolvasott vivoszövegbol

- Kivágás elott a felolvasás helyességét ellenorizni, hiba esetén a redundáns elem elovétele.
- Idobeli (esetleg spektrális) vizsgálat alapján határok megállapítása.
- Elemek elmentése az építoelem lista és a szabályrendszernek megfelelo logikus rendben (adatbázis, könyvtárstruktúra, stb.)

3.3.4.4. Egy lehetséges megvalósítási struktúra



Input number: "154."
(1) "_" "1" "100" "50" "4" "."
(2) Hungarian: "_" "100" "50" "4" "."
3 rules
German: "_" "1" "100" "4" "und" "50" "."
5 rules
(3) List of files to be concatenated

Block diagram of NTS algorithm implementation

- Nyelvfüggo és nyelvfüggetlen modulok és adatbázisok

3.3.5. A muködo rendszer tesztelése és javítása

- Analízis szintézissel módszer (analysis by synthesis).

Olaszy törvény (1984): Ahhoz, hogy egy beszédszintetizátor fejlesztéséhez hozzá tudjunk kezdeni, elso lépésben létre kell hozni egy beszédszintetizátort.

- A durvább hibák megszüntetése után vesszük észre a finomabb hibákat.
- Szint, idotartam/sebesség kiegyenlítés, hangsebészet.

3.4. Szövegfelolvasó rendszerek (TTS)

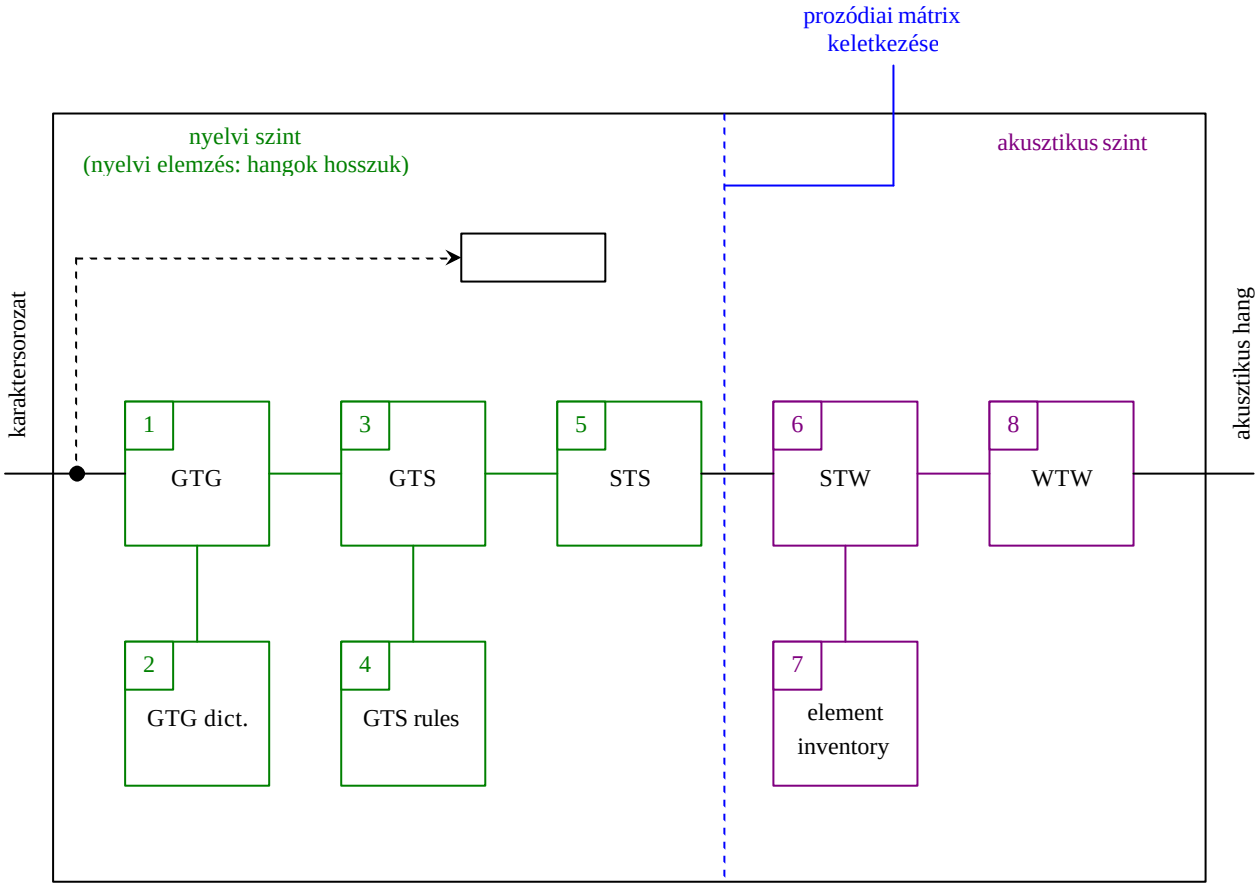
- **Szövegfelolvasó (text to speech):** adott nyelv köznapi szókincsében eloforduló szövegek felolvasása (kb. egy 8 éves gyerek szókincsének megfelelo)
- **Üzenet felolvasó (concept text to speech):** a kifejezni kívánt üzenetre vonatkozó jelekkel ellátott szöveg felolvasása
  - pl. [Conf_Req] A gépkocsi típusa [Car_Type] Volgswagen Golf
- **Többnyelvu TTS (miltilingual):** azonos építoelemek minél nagyobb halmazának egységes keretben történo felhasználása TTS rendszer megvalósításához több nyelven. Ideális esetben (ami cél és nem pedig a valóság) azonos program kód (és hardware), a nyelvfüggo adatok egységes szerkezetu, külso adatbázisban helyezkednek el.
- **Poligott TTS:** azonos hangon szóló TTS
  - zseniális paraméteres leírás, mely nincs emberi hanghoz kötve
  - egy bemondó sok nyelven mondja el a szöveget
- **Kötött tematikájkú (domain specific) TTS:** csak egy adott témaköru (pl. menetrend, idojárás, szállodafoglalás) szöveg felolvasására alkalmas rendszer. Átmenet egy hagyományos kötött szókészletu és egy TTS rendszer között.
- **képernyo felolvasó (screen reader):** számítógép monditor tartalmát értelmezo vakok és gyengénlátók számára. Nem tartalmaz TTS-t, csak illesztést képes alkalmazás és TTS között.

3.4.1. Osztályozási szempontok

- milyen nyelveken szeretnénk felolvastatni
- milyen szövegeket – egy teljes rendszert általában csak a TTS kimenete alapján ítélnek meg, a bemenetet nem látják.
  - szövegtípus: általános, szakszöveg, e-mail, SMS, stb.
  - mondattípus: kijelento, kérdo, felkiáltó, egyéb érzelem kifejezése, CTS
- milyen minoségben
  - érthetoség : intelligibility
  - természetesség: naturalness
    - és ezek nem is feltétlenül korrelálnak egymással
- milyen hangokon – egy illetve több hangon, amit kiemelünk, más hangon szóljon
- milyen paraméterek állíthatók
  - sebesség
  - hangmagasság
  - suttogás
  - rekedtség
  - szünetek hossza
  - betuzés
- milyen platformokon fusson
  - hardware
  - operációs rendszer (Windows, Unix, OS/2)
  - eroforrásigény, csatorna – nem mindegy, hogy mobiltelefonban vagy távközlési központban
- milyen vezérlési felületek, API-k
- bovítési, továbbfejlesztési lehetoségek – mit ad hozzá a felhasználó és mit a fejleszto, pl. rövidítésfeloldó
- milyen speciális igények merülnek fel – pl. IT, callback egy adott szó kimondásának elején/végén, kimondás állapotának lekérdezhetosége (menetrendnél
- milyen támogatást ad a TTS fejleszto az alkalmazásfejlesztonek

3.4.2. Felépítés: néhány alapprobléma

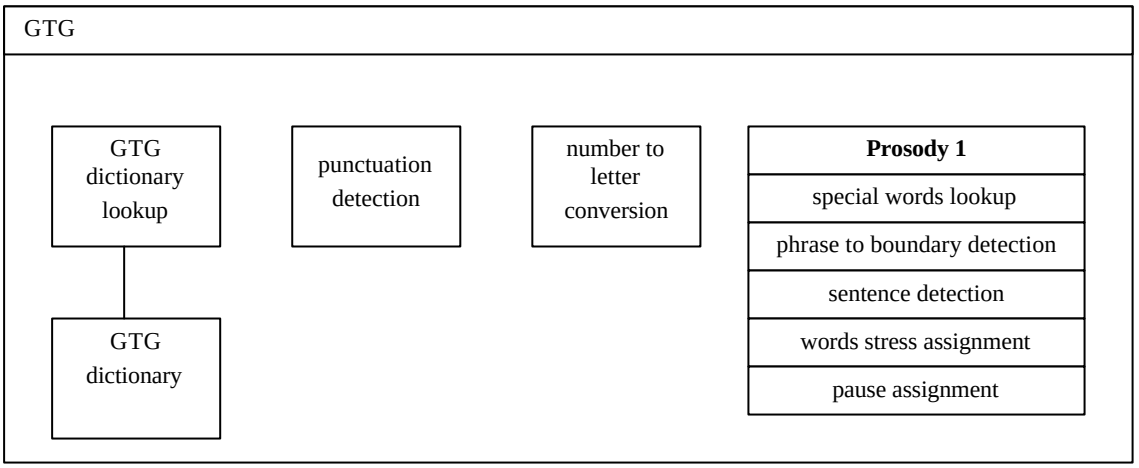
- Az írás diszkrét, a szavakat szünetek választják el. A beszédben a szavak folyamatosan következnek egymás után, csak nagyobb egységeket (prozódiai egység) választ el szünet. A beszédben a folyamatosság megértése teszi nehézé a megértést.
- Az írott hibákat másképp kezeljük: az akusztikus formára sokkal érzékenyebbek vagyunk.
- Fontos: a TTS bemenetére minél helyesebb és minél részletesebb szókimondást segíto információt tartalmazó jelsorozat érkezzen.



1. GTG : Grapheme to grapheme (írásjel→betu)
2. GTG dict. : GTG dictionary (szótár)
3. GTS : Grapheme to Sound (betu→hang)
4. GTS rules : szabály és szótár
5. STS : Sound to Sound (hang→hang)
6. STW : Sound to Wave (hang→hanghullám)
7. element inventory : hangelem-tár, akusztikai adatbázis
8. WTW : Wave to Wave (hanghullám-feldolgozás)

- 1-5 elvileg lehet nyelvfüggetlen, viszont 6-8 mindenképpen nyelvfüggo
- ha igazán általánossá akarjuk tenni, akkor nagyon bonyolult és nagy leíró nyelvre van szükség

3.4.2.1. GTG

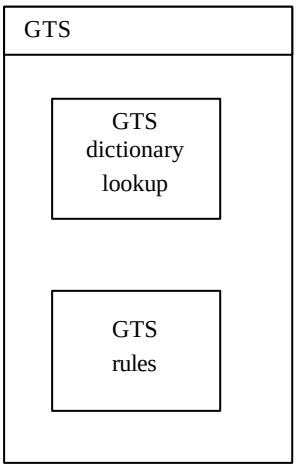


- punctuation detection: azért fontos, mert pl. egy mondatban pont sok helyen lehet (rövidítések, …, mondat vége, stb.)
- special words lookup: pontosvesszo, pont, vesszo, zárójel, csillag, bizonyos dolgokat nem mindig akarunk hallani
- phrase boundary detection: egységként kimondható szavak, frázisok között mikor tartunk szünetet
- sentence detection: intonáció egy mondatra

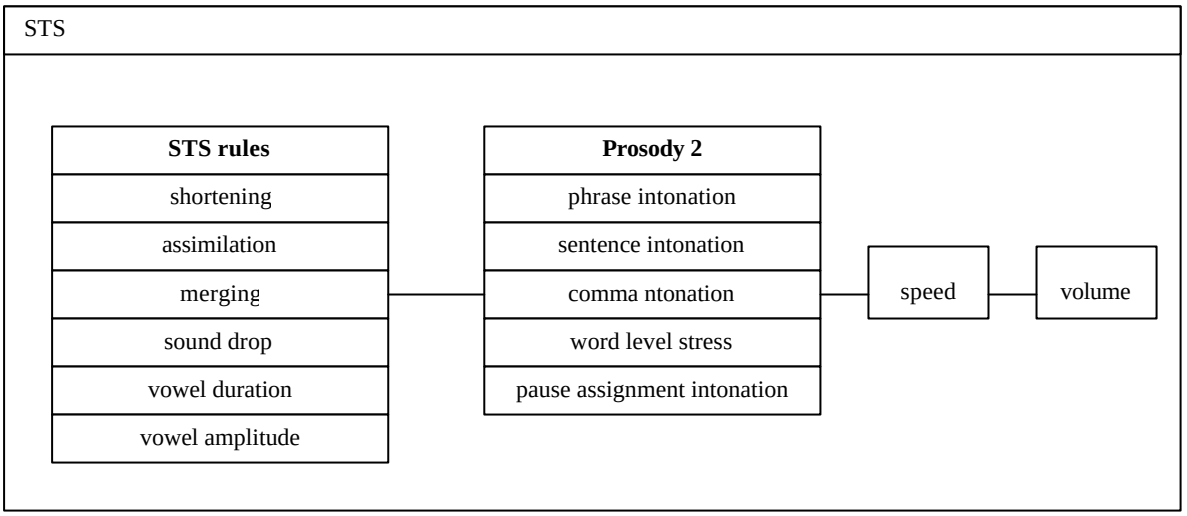
Példa: (6 soros idézet a BME szabályozási rendeletébol)
- szövegfelolvasó mit dolgozzon fel egy egységként (az egész egy mondat)
- többszintu nyelvi elemzés szükséges

Írott szövegnek nem egy az egyben felelnek meg a kimondott hangok
- hasonulások
- röviden írjuk, hosszan ejtjük és viszont
- mássalhangzó torlódások
- betukép helyes értelmezése szó illetve morféma határon (malacság, egészség)

3.4.2.2. GTS

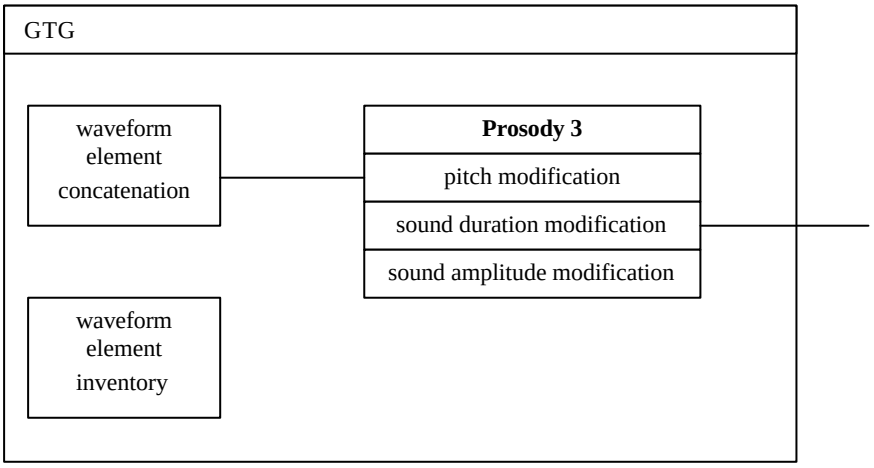


3.4.2.3. STS



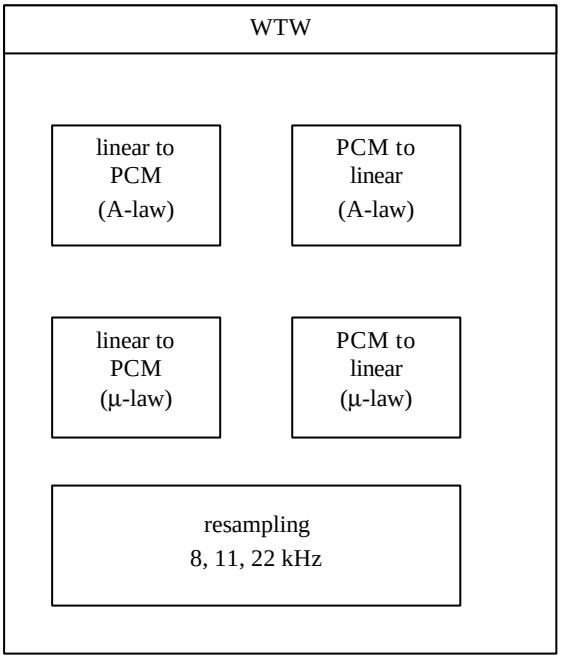
- speed: szünetek kivágása (vigyazni kell vele, mert ha figyelmetlenül vagdossuk ki a szüneteket, nem ugyanazt a hangot kapjuk)
- prosody 2: magasszintu leírás

3.4.2.4. STW



- waveform element inventory: valamilyen akusztikai adatbázis
  - paraméterek (pl. LPC) forráskódolása
  - hullámforma-kódolá
- prosody 3
  - a jó minoségu szövegfelolvasók esetén kulcsfontosságú
  - a prozódiai mátrixban eloírtakat el kell végezni (módosítások és folytonosság biztosítása)

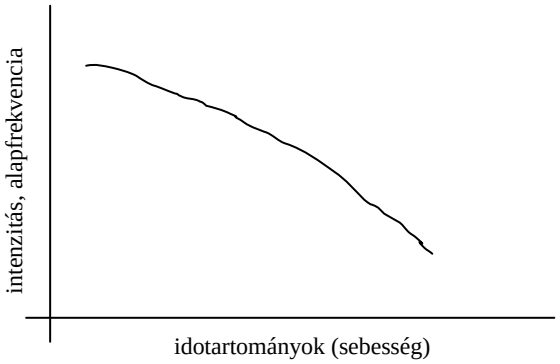
3.4.2.5. WTW



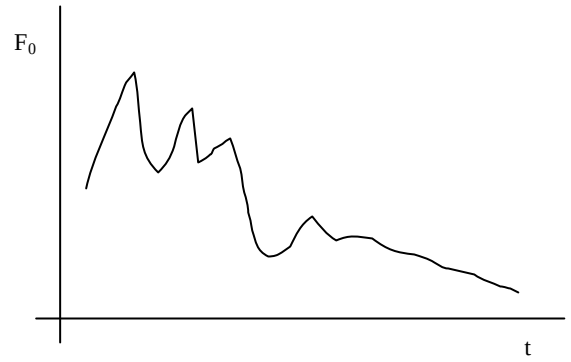
3.4.3. Néhány elvi probléma

- alapveto feldolgozási egység:
- ember esetében a mondatnál nagyobb
- mondat egy gépi megoldásban: a prozódiai algoritmusok legmagasabb szintje is a mondat

Pl. Kijelento mondat eso jellegu, ez azonban humán megközelítés, mérnöki módon hogyan fejezzük ezt ki



A prozódia három fo tulajdonsága
- intenzitás
- alapfrekvencia
- idotartománybeli jellemzok



nem folytonos: F0-t csak a zöngés hangoknál tudjuk értelmezni, felpattanó hangoknál pl. nem

- absztrakciós szintek:
- hang szint (legalacsonyabb)
- szótag szint
- szó szint
- tagmondat szint (prozódiai fázis)
- mondat
- ezekre mérheto, fizikai paramétereket kellene találni
- beszédfelismerés kulcsterülete: a sokrétu szinteket megkülönböztetni, elválasztani (lehetoleg minél jobban) – és a beszédben ezek folytonosak.

3.4.4. Megoldási stratégiák

3.4.4.1. Szabályalapú

- lebontás: címkézés
  - mondat: kijelento, kérdo, felkiáltó
  - szó: alany, állítmány, tárgy, határozó, jelzo; hangsúlyos/hangsúlytalan
  - szótag: hangsúlyos/hangsúlytalan
  - hang: szó eleje, szó közepe, szó vége; alacsony/mély hangrendu; magánhangzó/mássalhangzó; stb.
- szabályok megalkotása nagyon lassú – hiba esetén a szabályok kijavítása nehézkes
- a nyelv nem reguláris szerkezetu, vannak kivételek
- a nyelv változik, nem statikus

3.4.4.2. Gépi tanulás (macine learning)

- a gyakorlatból, mint nagy adatbázisból kinyerjük a szabályokat
- vegyünk sok, egymással összefüggésbe hozott, címkézett adatot – ez elég nagy adatbázis, címkézett szöveggel: neurális hálóval megvalósítva a rendszer következtetni tud
- a hosszú, absztrakciós szinten történo munkát kiváltjuk: sok adatban korrelációk, összefüggések keresése
- problémák
  - adatbázis létrehozása: több millió adatot kézzel kell felcímkézni (☹ hangszintig le kell menni)
  - a rendszer jól muködik arra az adatbázisra, amelyre be lett tanítva, de a többire nincs garancia
  - milyen alapon ítéljük meg egy rendszer jóságát (Ezt a felhasználó dönti el!)

Az ember többnyire rosszul turi a minoség változását. Ha megszoktunk egy adott minoséget (még ha rossz is), nehezen viseljük, ha az megváltozik. Ebben van a szabályalapú rendszerek egyik nagy elonye: kiszámíthatóság.

A valóságban a két módszer ötvözetét használják. Egy adott, zárt problémakört fed le a gépi tanulás módszere, a kimaradó halmazra valamilyen szabályalapú megoldást alkalmaznak.

3.4.5. TTS tervezése: a hanganyag (akusztikai elmebázis)

- az akusztikai elembázis nagyjából megfelel egy adott nyelv betukészletének (hangjainak)
- fonéma – graféma konverzió problémái:
- a fonéma minimálhalmaz, diszkrét elemek definíciója, a természetes beszéd azonban folytonos
- hangelemek (hangkód) – a fonémák kibovítése, pl. tájszólások e-jei, ng, rövid/hosszú magánhangzók
- ezek átmeneti tulajdonságait is figyelembe kell venni: megoldási ölet

3.4.5.1. Diád (diphone)

- pl. legyenek az adott hangsorozat hangkódjai: 1,5,11,13. Ekkor vesszük a 1-t és az 5-t a közepéig, aztán az 5-t a közepétol a 11 közepéig, majd a 11-t a közepétol és a 13-mat.
- Minden elemet az alábbi módon definiálunk:

elemhatár – hanghatár – elemhatár

	1	2	3	4	5		11	12	13	…	50
1					x						
2											
3											
4											
5							x				
⋮											
11									x		
⋮											
50											

- Így az elem tartalmazza mindegyik hang bizonyos stabil szakaszát és az átmenetet.
- ez az ún. diád (diphone), ezzel csak az a probléma, hogy négyzetes a fonémák számával

Diád peremfeltétel: diádhatáron a hangokban legyen folytonos átmenet (bizonyos esetekben megoldható: szünet + s; bizonyos esetben nem oldható meg: rövid magánhangzó, több hatás, elöl is és hátul is hat)

3.4.5.2. Triád (triphone)

- vegyünk hármas egységeket, a problémás elem legyen középen, az elemhatárt tegyük oda, ahol a vágás gond nélkül megteheto

elemhatár + 2 hanghatár

- a fonémák köbével arányos

3.4.5.3. További kiterjesztés

- azon elemeket, amelyek gyakran elofordulnak, tároljuk diádos/triádos elemekbol
- változó méretu adatbázis (non-uniform database)

3.4.5.4. Adatbázis-elemek ábrázolása

- alapvetoen aszerint, hogy a prozódiai módosításokat el akarjuk-e rajtuk végezni vagy sem, és ha igen, akkor milyen technikával
  - hullámforma: módosítás nélküliek
  - parametrikus forráskódolt: formáns, LPC

3.5. PSOLA (Pitch Synchronous OverLap Add) algoritmus

(zöngeszinkron átlapoló-összeadó)

3.5.1. Alapötlet

- tekintsük s(n)-t, mint egy FIR szurot egy i(n) impulzussorozatra
- ekkor szét kell szedni a jelet valamilyen pozícióban megjeleno impulzussorozatra és az azokra adott elemi válaszok összegére

$i(n) = \sum_{k=-\infty}^{\infty} d(n - P_a(k))$ és a válaszfüggvény: $h^{P_a(k)}(n) = s(n) \cdot w^{P_a(k)}(n)$

- w: valamilyen ablakot elhelyezünk a hullámformán Legyen ez az ablakfüggvény:

$$^{P_{a(k)}}(n) = \left\{ \begin{array}{c} 0, \quad ha\ n < P_a(k-1)\ vagy\ n > P_a(k+1) \\ 0.5 - 0.5\cos\left(\frac{p(n - P_a(k-1))}{P_a(k) - P_a(k-1)}\right), \quad ha\ P_a(k-1) < n < P_a(k) \\ 0.5 - 0.5\cos\left(\frac{p + p(n - P_a(k))}{P_a(k+1) - P_a(k)}\right), \quad ha\ P_a(k) < n < P_a(k+1) \end{array} \right\}$$

- a függvény érdekessége, hogy nem kauzális

3.5.2. Szintézis:

$$\hat{s}(n) = \sum_{n=-\infty}^{\infty} i(m) \cdot h^m(n) \quad és \quad i(m) = \sum_{k=-\infty}^{\infty} d(m - P_a(k))$$

$$\hat{s}(n) = \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} d(m - P_a(k) h^m(n)) = \sum_{k=-\infty}^{\infty} h^{P_a(k)}(n) = s(n) \cdot \left(\sum_{k=-\infty}^{\infty} w^{P_a(k)}(n) = s(n) \right)$$

- Ha egyetlen ilyen létrejött az adatbázisban (megvannak a válaszfüggvények)
- alapfrekvencia változtatás: egy pitchmark sorozat változtatás: ha növelni szeretnénk az alapfrekvenciát, akkor közelebb hozzuk oket, így a válaszokat összeadogatva nagyobb lesz a frekvencia
  - ☹ nagyon közel nem hozhatjuk oket, mert akkor egymásra csúsznak
  - ☹ nagyon távol sem vihetjük oket, mert akkor a jel elhal két impulzus között ,és ott csend lesz
- Ezzel az algoritmussal kb. 20%-ot lehet változtatni az alapfrekvencián.

4. BESZÉDFELISMEROK

4.1. Bevezetés

4.1.1. Felismerési feladatok

- a gépi rendszer ismerje fel a beszédet
- beszéddetekció: annak felismerése, hogy beszéd van vagy nincs (sokszor része a beszédfelismeroknek, de önmagában is hasznos lehet[1])
- zöngés/zöngétlen meghatározás – leginkább csak támogatja a beszédfelismeroket, de néha önálló feladatnak is tekintik
- beszélo felismerés
- speciális esete a beszélo azonosítás
  - hagyományos módszer mintaszöveg felvétele, amit belépéskor el kell mondani
  - véletlenszeruen kisorsolt minta: sok mintát vesznek fel, és ezek közül egyet véletlenszeruen sorsol a rendszer a belépéskor

4.1.2. Beszédfelismerés osztályozása

- kis (kötött) szótáras, kb. 100 szó ⇔ nagy szótáras (kötetlen szótáras), 20-80000 szó[2]
- személyfüggo ⇔ személyfüggetlen
  - személyfüggo: egy személy beszédét ismeri fel, általában adaptív rendszer, egy adott személyre rátanul
  - személyfüggetlen: nagyon sok mintával dolgozik, a személyfüggoséget megpróbálja kiátlagolni
- izolált szavas ⇔ kapcsolt szavas ⇔ folyamatos beszéd
  - izolált szavas: egymástól hosszú idovel elválasztott szavak (pl. utasítások)
  - kapcsolt szavas: a szavak közti szünetek minimálisak
  - folyamatos beszéd: diktáló rendszerek
- jó minoségu beszédbol felismerok ⇔ robosztus rendszerek
- jó minoségu beszédbol felismerok: mindig innen indiuul a felismerés, és valamilyen trükkel sikerül robosztussá tenni
- robosztus rendszerek: elég nagy zaj mellett is felismerik a beszédet[3]

4.1.3. Filozófiai probléma

- 1965-ig az a tendencia volt, hogy próbáljuk meg leutánozni, hogy mi történik az agyban a beszédfelismerés során
- 1965-ben új elmélet: az agyban született mondanivaló kódolása a beszéd, tehát a szájon keresztül kijövo hangból kell tudnunk következtetni a mondanivalóra
- nehézség: egyazon mondanivalónak végtelen sok reprezentációja van akusztikai szinten
  - nyelvi szinten ugyanazon dolgok akusztikai megjelenése különbözo
  - a végtelen sokféle reprezentációk közül melyek azok, melyek 1 fonémának/átmenetnek tekinthetok és melyek nem

[1] Pl. zajos környezetben segélykiáltás detektálása

[2] Bizonyos nyelveken ez már gyakorlatilag diktáló rendszernek (STT: Speech to Text) tekintheto. Magyarban ez nem így van, mivel a magyar toldalékozó nyelv. Így egy adott szó felvétele nem jelenti a szó összes megjelenési formájának felvételét. Erre illusztratív példa, hogy míg az angolban (amely nem egy toldalékoló nyelv) egy 20000 szavas felismero jó minoségu, addig egy ugyanennyi szót tartalmaz magyar nyelvu felismero kb. az idojárás jelentést tudná viszonylag nagy biztonsággal felismerni.

[3] Pl. harci helikopterekben használnak beszédfelismeroket bizonyos parancsokra (izolált szavas, de nagyon robosztus)

4.1.4. A beszédfelismerés három komponense

- lényegkiemelés (feature abstraction)
  - a hanghullám változásaiból olyan elemeket próbálunk kiemelni, melyeknek kicsi az intrindividuális és az interindividuális jellemzoje (függetlenül attól, hogy ki mondta, milyen érzelmi állapotban mondta)
  - redukáljuk az adatmennyiséget (kb. 1/10 részére csökkentheto)
- mintaillesztés
  - ugyanazt a szót nem lehet kétszer ugyanabban a ritmusban kimondani, ezért a mintaillesztés legfontosabb feladata a különbözo ritmikájú kiejtések közötti ritmuskülönbségek kiküszöbölése
- utó/elofeldolgozás
  - utófeldolgozás: biztosabbá teszi a felismerést
  - elofeldolgozás: könnyebbé teszi a felismerést

4.2. Lényegkiemelés

- az idofüggvényt keretekre (ablakokra) bontjuk, ezek 10-30 ms hosszú ablakok
- az ablakokat 50%-os fedésben helyezzük egymásra

4.2.1. Az ablak alakja

4.2.1.1. Rectangular (négyszögletes)

4.2.2. Hamming ablak

Ha a beszéd idofüggvénye f(t) és az ablak idofüggvénye w(t), akkor a kiablakolt függvény a(t)=f(t)·w(t). A spektrális jellemzésre igen jó a Fourier transzformáció:

$$A(w) = F(w) * W(w) = \int_{-\infty}^{\infty} F(a)W(w_0 - a)da$$

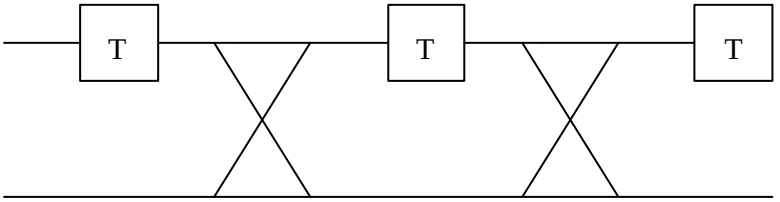
Olyan ablakfüggvényt kell választanunk, amelynél a kiablakolt beszéd spektrálisan legjobban hasonlít az eredetire, azaz az ablakfüggvény legjobban hasonlít a dirac függvényre: $\delta(t)$. A Hamming ablak bizonyos értelemben jobban hasonlít a dirac függvényre

☺ ω_0-tól távol nagyon gyorsan lecseng, kevés összetevot vesz figyelembe

☹ a közvetlen ω_0 melletti összetevok jobban befolyásolják a kiablakolt jelet[1]

4.2.3. Lényegi jellemzok

- szegmens-energia
- F_0, amennyiben a hang zöngés
- nullátmenetek száma a szegmensen belül (beszéddetekciónál fontos)
- LPC paraméterek (w_i)
- Gráfstruktúrában (rácsszerkezetu gráffal) meg lehet határozni a kódolót és a dekódolót is



[1] Az a sáv amelyen belül simít, pszichoakusztikailag nem zavaró, mert a fül is simít, és ez a kritikus sávon nem nyúlik túl

- $k_i \le 1$ – ezek az ún. parcor együtthatók, melyek egyre több korrelációt vonnak kis a jelbol, végül teljesen korrelálatlan lesz
- Más megoldás, ha valaki leírja a vokális traktust, mint egy változó keresztmetszetu csövet, ekkor kapjuk az ún. área együtthatókat.
- $r_i = \frac{A_{i+1} - A_i}{A_{i+1} + A_i}$, és érdekes módon $r_i = -k_i$
- A beszédhullámokból a beszédkeltore lehet következtetni, amikor a parcor együtthatókat határozzuk meg.

4.2.3.1. Spektrális jellemzok

- általában minták állnak rendelkezésre, ezért DFT-t (diszkrét Fourier transzformációt) alkalmazunk.
- a DFT a halmozott spektrum mintáit adja (számsorozat DFT-je a számsorozathoz tartozó spektrum elégséges mintáit adja)
- 1965-ben Cooley és Tuckey felfedezték az FFT-t (fast Fourier transform)[1]

4.2.3.2. Kepsztrális jellemzok

1962-ben Bogert, Healy és Tuckey észrevette, hogy a beszédet nagyon jól jellemzi a kepsztruma, amelyet az alábbi módon definiáltak:

$$c(q) = \left| \Im^{-1} \left\{ \log(\left| \Im\{a(t)\} \right|^2) \right\} \right|^l \quad ahol\ l = 1,2$$

- c(q) a kepsztrum, ido jellegu mennyiség (de nem ido)
- konvolvált idofüggvények kepsztruma
  - természetes beszéd: gerjesztés $*$ vokális traktus súlyfüggvénye
  - rögzített beszéd: beszéd $*$ mikrofon súlyfüggvénye
- ha sikerül különválasztani a kepsztrum segítségével a vokális traktus és a beszéd súlyfüggvényét, akkor a beszéd fobb jellemzoit különválaszthatjuk.

4.2.3.3. Kepsztrum analízis

$$c(q) = \left| \Im^{-1} \left\{ \log(\left| \Im\{a(t)\} \right|^2) \right\} \right|^l \quad ahol\ l = 1,2$$

- a(t) pedig a kiablakolt idofüggvény: a(t)=f(t)w(t)
- a logaritmálás lényege, hogy ha a(t) egy eleme nem szorzat, hanem konvolúvió, akkor össze lehet oket adni.
- az inverz Fourier transzformáció hatására ido jellegu térbe transzformáljuk a jelet, melynek változója a kefrencia.

Legyen a(t) = x(t) $*$ y(t), ahol x(t) a zönge, y(t) pedig a vokális traktus súlyfüggvénye

$$\Im\{a(t)\} = A(w) = X(w)Y(w) \qquad /^2, \log$$

$$\log \left|A(w)\right|^2 = \log \left|X(w)\right|^2 + \log \left|Y(w)\right|^2$$

$$g_a(q) = g_x(q) + g_y(q)$$

$$C_f(q) = \left(g_x(q) + g_y(q)\right) \cdot \left(\bar{g}_x(q) + \bar{g}_y(q)\right) = \left|g_x(q)\right|^2 + \left|g_y(q)\right|^2 + g_x(q)\bar{g}_y(q) + g_y(q)\bar{g}_x(q)$$

- az elso két tag pedig $c_x(q)$ és $c_y(q)$, azaz a zönge és a vokális traktus kepsztruma

[1] Az FFT DFT esetén egy n elemu számsorozat, és míg DFT-hez n^2 szorzás és n^2 komplex összeadást kell végezni, míg az FFT-hez $\frac{n \cdot \log_2 n}{2}$ szorzást és $\log_2 n$ összeadás szükséges.

Ha x kepsztruma valamilyen kefrencia tartományban domináns, és y egy másikban, akkor a kettejük szorzata éppen 0-t ad. Ha a tartók elkülönülok, akkor a kepsztrum a két jel kepszrtumainak összege. Ilyen esetben a kepsztrum analízis dekompozícióra alkalmazható.

- Lényegkiemelést alkalmazva keretenként egy paramétervektort kapunk
- 10/20-adfokú LPC analízist használva 10/20-adfokú kepsztrumot kapunk

A paramétervektort az alábbi módon származtatjuk

$$p_t = \begin{bmatrix} p_i^t \quad i = 1..l & \text{primer paraméterek} \\ p_i^t - p_i^{t-1} & \text{delta paraméterek} \\ \sum_{j=-2}^{2} a_j \cdot p_i^{t+j} & \text{delta paraméterek lineáris kmbinációj a} \end{bmatrix}$$

- sokáig azzal próbálkoztak, hogy a paraméterek számát növelték, később azonban kiderült, hogy bizonyos paraméterszám után ez nem ad többjellemzot
- egy szegmensbe ne csak a saját paramétereit vegyük bele, hanem az elozo szegmensbol is néhány paramétert
- a beszédet az a változás is jellemzi, hogy mi kerül a következo szegmensbe
- a p_i^t p_i^{t+1} és p_i^{t-1} segítségével lineáris kombinációval megkapható
- a mai kísérletek nagy rész arra irányul, hogy a vektorparamétereket milyen súllyal vegyük figyelembe ahhoz, hogy optimális legyen

4.3. Mintaillesztés alapjai

Adottak pl. izolált szavak (lényegkiemelt vektorsorozatokkal), kefrenciák, prototípusok (r_k vektorsorozatok) illetve fonémák (ezekhez is vektorsorozatok tartoznak), vagyis a felismerés alapjául szolgáló nyelvi egységet reprezentáló vektorsorozatok. Feladat, hogy az ún. tesztkiejtésbol (felismerheto kiejtés) meg tudjuk állapítani, hogy melyik referenciához hasonlít a legjobban.

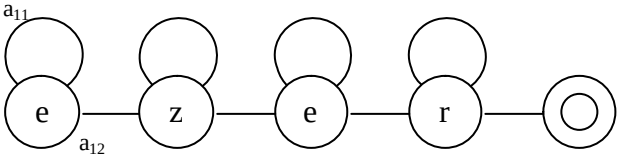
Legnagyobb probléma, hogy ugyanazt a szót az emberek különbözo ritmusban képesek kiejteni, de ugyanez igaz egy embernél ugyanazon szó kétszeri kiejtésénél. Meg kell tehát találnunk az a technikát, amivel a megfelelo dolgok lesznek összeillesztve.

Erre három módszer létezik, ebbol kettonek statisztikai megfigyelés az alapja, a harmadik sablon (template) alapú.

- HMM – Hidden Markov Model (statisztikus)
- ANN – Artificial Neural Network (statisztikus)
- DTW – Dynamic Time Warping (sablon alapú)

4.3.1. Mintaillesztés HMM-mel

Ritmikai változások figyelembevétele pl. izolált szavas beszédfelismerésnél:



A modell: lépni kényszerül minden 10. ms-ban, de nem kényszerül ellépni onnan. Így ezzel a technikával alapvetoen ki lehet küszöbölni az alapveto ritmusbeli különbségeket. Lehetséges ugró él is, ha valamelyik hangot nem ejtjük ki.

Miért hívják ezt rejtett Markov modellnek? Azért, mert a véges automatáknál megszokott módtól eltéroen itt nem tudjuk, hogy a folyamat milyen állapotban van. Erre a megfigyelésbol kell következtetnünk.

A modell kiad P_T vektort, miközben az állapotokban eljut az N. állapotig. A megfigyelési sorozatot O-val jelöljük (mint observation). Aközben az emisszióból (azon vektorok, amelyeket a Markov folyamat emittál) nem tudjuk megállapítani, hogy melyik állapotban vagyunk.

Egy állapothoz sokféle vektor tartozhat, ezért inkább valószínuségekkel számolunk.

$P(O_t \mid q_j)$ egy valószínuségre jellemzo érték, ahol q_j a j-edik állapot O_t pedig folytonos értékkészletu. Minden egyes ponthoz suruségfüggvény-értéket rendelünk. Ezt a suruségfüggvényt adatbázisokból kell meghatározni. Az egyszeruség kedvéért csak a Gauss-eloszlások lineáris kombinációit tekintjük.

$b_j(O_t) = \sum_{m=1}^{M} C_{jm} \cdot G(O_t, m_{jm}, s_{jm})$, ahol G(x, μ, σ) egy μ várható értéku σ szórású Gauss eloszlás.

Szemléletesen ennek az a jelentése, hogy a j. állapotban mennyi a valószínusége, hogy O_t megfigyelés adódik. Az adatbázisok alapján meg kell határoznunk az átmeneti valószínuségeket. A tanítás során:

- e_1 hozzárendeljük az elso Markov folyamatot minden felismerheto elemhez
- e_j $b_j(O_t)$ j=1,2…N, $\{a_{ij}\}$

Végül minden ei-hez tartozik egy $b_j(O_t)$ készlet.

Felismerés: érkezik egy tényleges megfigyelést tartalmazó vektor, ekkor ki kell keresni egy olyan Markov láncot, amely a legnagyobb valószínuséggel tudja ezt követni. Meg kell nézni, hogy az egyes Markov modellek milyen valószínuséggel képesek ezt az O_t –t produkálni, és a legnagyobb valószínuségu modell által leírt nyelvi elemet tekintjük felismernek.

A Baum-Welch algoritmus egyértelmuen meghatározza a c, μ, σ-t, de adott N esetében. A Viterbi algoritmus különbözo Markov modellekre is alkalmazható.

- feltételezve O_t –t létezik egy legnagyobb valószínuségu (optimális) út a kiindulástól i-be
- $\delta_t(i)$ legyen ezen út valószínusége
- így minden pontra meghatározható az odavezeto optimális út valószínusége
  1. inicializálás: $d_0(i) = 1$
  2. indukciós lépés: $d_{t+1}(j) = \max_{\forall i}\{d_t(i) \cdot a_{ij}\} \cdot b_j(O_{t+1})$, és *i = 1 .. N, j = 1 .. N*
  3. leállás feltétele: *t = T*
  4. kimenet: $\max_i\{d_T(i)\}$ és *i = 1 .. N*
- végül megkapjuk, hogy melyik az a Markov modell, amelyik a legvalószínubb.
- izolált szavak esetén a modellhez egyértelmuen tartozik a felismert szó
- Vannak olyan esetek, amikor az is fontos, hogy mi volt az optimális út, nem elég a célállapot. Ilyenkor minden pontba azt az értéket írjuk, amelyik állapotból jöttünk, azaz $\arg\max_i\{d_t(i) \cdot a_{ij}\}$-t. Így T idopontból visszafelé az optimális út meghatározható.

4.3.2. Dinamikus idovetemítés (Dynamic Time Warping)

- pl. izolált szavas felismerésénél, itt ugyanis nincs lehetoség statisztikus felismerésre
- felvesszük a felismerendo szöveget és a referenciamintára próbálunk illeszteni
- a legjobban hasonlító minták a legkisebb távolságra vannak a referenciától – definiálni kell tehát egy d távolságot
- Legyen
  - $\vec{t}$ tesztvektor
  - $\vec{r}_k$ referenciavektor, k= 1..M
  - és így $\arg\min_{\vec{r}_k}\{d(\vec{r}_k,\vec{t})\}$

Az elobb említett távolságot az alábbi módokon definiálhatjuk (általában az elso kettot szoktuk használni):

$$d(\vec{r}_k,\vec{t})=\left\{\begin{array}{c}\sum_{i=1}^{M}(r_k(i)-t(i))^2\\ \sum_{i=1}^{M}|r_k(i)-t(i)|\\ \ldots\\ \max_i|r_k(i)-t(i)|\end{array}\right\}$$

Probléma akkor van, ha $\vec{t}$ és $\vec{r}_k$ vektorok különbözo hosszúak. Ez abból adódhat, hogy egyrészt a különbözo bemondások különbözo sebességgel történnek, másrészt ugyanazon bemondáson belül sebességingadozás is felléphet.

Az egyszeru lineáris vetemítés akkor lenne jó, ha kizárólag sebességkülönbségrol van szó (pl. az ember is nyávogna, ha gyorsabban beszél, mint a magnóra felvett szöveg, ha gyorsabban játsszuk le).

Ezért nemlineáris vetemítésre van szükség.

- legyen $F_i(t)$ egy vetemítogörbe
- ezek kijelölnek bizonyos tartományt a $\vec{t}$ - $\vec{r}$ síkon (bizonyos irányokra majd nem is lesz szükség, úgyhogy ezek majd bizonyos korlátokat szabnak)
- a $d(\vec{r},\vec{t})=\min_{F_i}\{d(r(t),t(F_i))\}$ távolságot úgy kell meghatározni, hogy az adott vetemítogörbére vett távolság a legkisebb legyen
- a vetemítogörbe tulajdonságai
  - monoton növekszik
  - lokális korlátok (extrém körülmények között túllépheto)
  - teljes optimum: lokális optimumokon keresztül valósul meg[1]

[1] Bárhol megállva a vetemítogörbén, addig a pontig vezeto út optimális kell, hogy legyen

Példa

$\vec{t}$ = {2,6,8,9,8,3}
$\vec{r}$ = {1,6,9,6,5}
d:= $\sum_i |r(i) - t(i)|$

1. lokális tulajdonságok kiszámítása

r_6	2	3	6	3	2
r_5	7	2	1	2	3
r_4	8	3	0	3	4
r_3	7	2	1	2	3
r_2	5	0	3	0	1
r_1	1	4	7	4	3
	t_1	t_2	t_3	t_4	t_5

2. akkumulált tulajdonságok kiszámítása

r_6	30	11	9	6	**6**
r_5	28	8	3	**4**	7
r_4	21	6	**2**	5	8
r_3	13	3	**2**	4	7
r_2	6	**1**	4	4	5
r_1	**1**	5	12	16	19
	t_1	t_2	t_3	t_4	t_5

Az eljárás elonyei
☺ gyorsan tanítható, gyakorlatilag 1 mintával is valamilyen szinten muködhet
☺ a minták a vetemítofüggvény mentén átlagolhatók

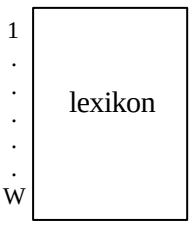
4.3.3. Nyelvi modell integrálása HMM illesztéséhez

ha nem csak izolált szavakat akarunk használni

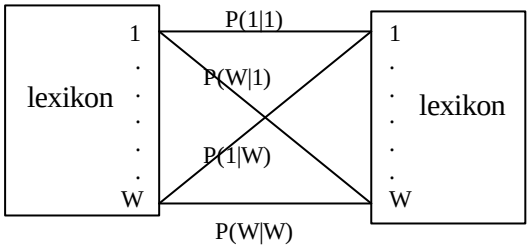
Alapötlet: az izolált szavas HMM Markov modelljeit összevonjuk, így s rendszerbe nyelvtani információt viszünk a modell topológián keresztül
☹ a természetes nyelvek nem írhatók le determinisztikus nyelvtannal
☹ méret is probléma: sztochasztikus n-gram modellek (n=1,2,3): milyen valószínuséggel fordul elo n adott szó egymás után
☹ a szöveg nem ugyanaz, mint a hangsor, tehát kiejtésmodellezés is szükséges

Legyen W a szótárméret

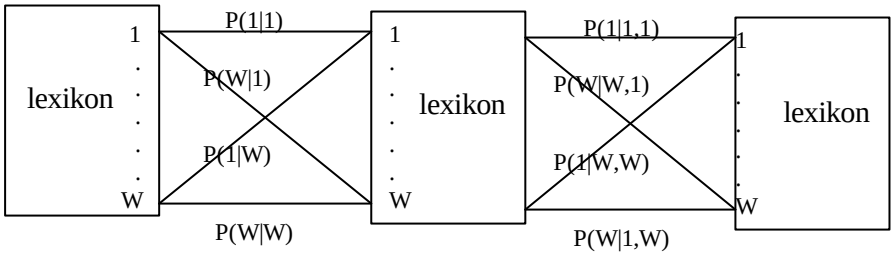
1-gram modell



bigram modell



trigram modell



Felmerül az optimalizálási igény: N = 20000, 20 modellállapot (szó), akkor a HMM állapotok szám: 20000*3*20 = 1.2 millió, súlyok: W^3-bel arányosak.

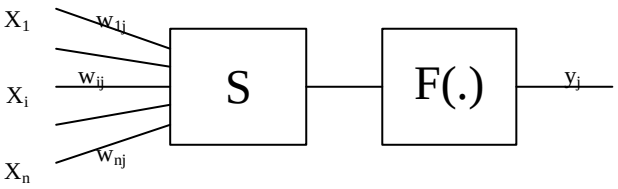
Keresés optimalizálása:
- keresési igény megadása (beam search)
- token passing (útvonal + útvonal valószínusége), teljesen értelmetlen irányok elhagyása

4.3.4. Mesterséges neurális háló (ANN)

izolált szavas beszédfelismerésre alkalmas, arra bevált

4.3.4.1. Mesterséges neuron

A j-edik neuron

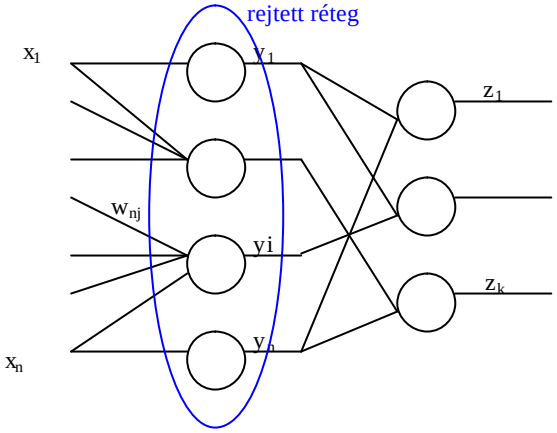


w_{ij}: a j-edik neuron i-edik bementének súlyvektora (weigthing factor)

$y_j = F\left(\sum_{i=1}^{n} w_{ij} \cdot x_i + \Theta_j\right) = F(a)$, ahol Θ_j a j. neuronra vonatkozó küszöbérték.

4.3.4.2. Egyszerusített Multi-Layer Percepton, 1 rejtett réteg

Minden bemenet hathat akármelyik azt követo neuronsorozat bármely elemére, és minden kimenet hathat az azt követo neuronsorozat bármely elemére.



Az általános neuronhálótól abban különbözik, hogy a neuronok rétegekbe rendezettek, nem tetszolegesen összekötöttek, nincs visszaút, egy réteg mindig csak egy magasabb rétegbe mehet, tehát a neuronhálót, mint irányított gráfot tekintve, ez a gráf DAG. Ezenkívül minden rétegben ugyanazt az F függvényt alkalmazzák (beszédfeldolgozásban rétegenként nem alkalmaznak külön függvényt, hanem az egész hálóban 1 F függvényt használnak)

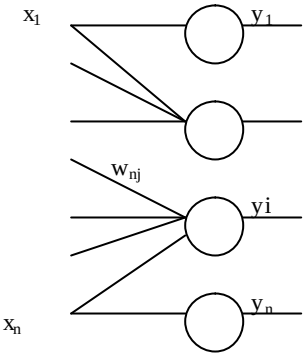
4.3.4.3. Beszédfelismerés ANN-nel: példák

1. x = (x1, x2 …xN) legyen az egy kerethez tartozó tulajdonságvektor: zk ≈ 1, ha x abból a hangból származik, amit mi a k-adik sorszámmal láttunk el, minden más esetben egy 0-hoz közeli érték. Így keretenként fonémákat akarunk felismerni
2. x = (x1, x2 …xl | xl+1, xl+2 …x2l | x2l+1, x2l+2 …x3l | … | x4l+1, x4l+2 …x5l), N = 5l, ezek pedig a t-2, t-1, t, t+1, t+2 keretekbol származó tulajdonságvektorok: zk ≈ 1, ha ez olyan szótagból származik, amelynek közepén t fonéma van, minden más esetben 0. Az ilyen perceptont idokésleltetéses perceptonnak nevezzük.
3. sok bemenetet megengedve (pl. izolált szavas számfelismero): 30-40 keretnyi bemenetet engedünk meg pl. 1000 neuron bementre, ez 10 kimeneten jön ki. A rejtett rétegben kb. 60-70 neuron található. Ez igen gyakori alkalmazása a MLP-nak.
4. elofeldolgozásra is használják: a fonémákat osztályokba sorolják (magánhangzó, félmagánhangzó, afrikáta, felpattanó, zárrés), és ezzel a szegmentáláson kívül a szegmensekhez a fonetikai osztályt is hozzárendeli. Ha az elotte/utána lévot is megmondja, akkor címkéz. A címkézés finomságát lehet változtatni. Tehát izolált szavakat, szótagokat NN-nel fel lehet ismerni, de magasabb szintu elemzéshez elokészítésnek is alkalmazható.

4.3.4.4. A muködés lényege: w_{ij}-k és F helyes megválasztása

- ezekre nem ismerünk analitikus módszert
- manapság w_{ij}-ket tanítással határozzuk meg
- nyelvi elemekkel bombázzuk a neurális hálót, és w_{ij}-ket úgy változtatgatjuk, hogy jól muködjön
- matematikailag bizonyítható, hogy ha az N dimenziós vektor olyan osztályokból származik, mely osztályok az N dimenziós térben – akár nem lineáris területtel is – elválnak, akkor a kimenet meghatározható.

4.3.4.5. Tanítás szemléltetése egyrétegu, nemlinearitás-mentes hálózatokra



x^P – a P-edik tanításkor bemeno sorozat
y^P – a P-edik tanításkor megjeleno tényleges kimenet
t^P – a tanítás során elvárt kimenet (target)

$y_j^P = \sum_i w_{ij} x_i^P$, ehhez kell definiálnunk egy hibát és egy súlytényezot, amivel korrigálni fogunk.

A hiba: $E = \sum_P \left[\frac{1}{2} \cdot \sum_j \left(t_j^P - y_j^P \right)^2 \right]$ $E^P = \frac{1}{2} \cdot \sum_j \left(t_j^P - y_j^P \right)^2$, E^P pedig a P-edik bemondással kapcsolatos hiba.

Optimalizálandó faktorok az y-okban lévo w_{ij}-k. Gradiens módszert alkalmazva a súlytényezok: $\Delta w_{ij} = h\left(-\frac{\partial E}{\partial w_{ij}} \right)$, ahol η egy alkalmasan választott konstans 0.02 és 0.1 között.[1]

$$\Delta w_{ij} = h\left(-\frac{\partial E}{\partial w_{ij}} \right) = h\left(-\frac{\partial}{\partial w_{ij}} \sum_P E^P \right) = \sum_P h\left(-\frac{\partial}{\partial w_{ij}} E^P \right) = \sum \Delta w_{ij}^P$$

mivel

$$\frac{\partial E^P}{\partial w_{ij}} = \frac{\partial E^P}{\partial y_j^P} \cdot \frac{\partial y_j^P}{\partial w_{ij}},$$

hiszen y_j^P az w_{ij} függvénye. Alkalmazva a parciális deriválás tulajdonságait: E^P parciális deriváltja csak az adott j esetében nem lesz 0.

Tehát a $t_j^P - y_j^P = -d_j^P$, y_j^P parciális deriváltja pedig éppen i=j esetben nem 0. Vagyis

$$\Delta w_{ij}^P = h \cdot d_j^P x_i^P$$

azaz

$$\Delta w_{ij} = \sum_P w_{ij}^P = h \sum_P d_j^P \cdot x_i^P .$$

[1] Ha η túl nagy, akkor oszcillál a kívánt érték körül, ha túl kicsi, akkor pedig nagyon lassan konvergál a kívánt értékhez.

4.4. Beszélofelismerés

4.4.1. Bevezetés

4.4.1.1. Ki volt a beszélo?

- megállapítható-e az elhangzó beszéd alapján a beszélo személye ha ismerjük az illetot, illetve ha nem

4.4.1.2. Feltételezés

- az agyunkban létrejövo neurális spektrogram tartalmazza a beszélo ismérveit
- vajon ez az információ, ezek a paraméterek olyan mértékben jellemzoek-e a beszélore, mint pl. az ujjlenyomat?

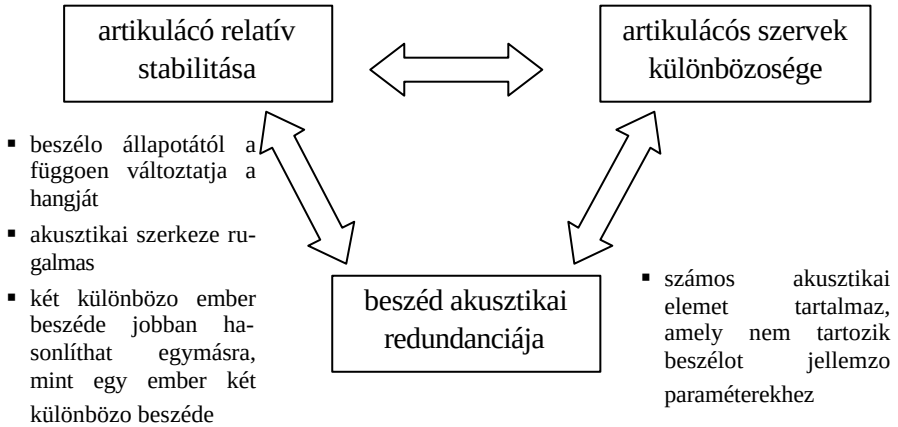
4.4.1.3. Alkalmazás

- a beszélo személy vizuális megjelenése nem fontos (belépteto rendszerek, információs rendszerek, telefonos lekérdezo rendszerek)
- kötött szótáras esetben létezik 100%-os felismerés, itt a beszélonek érdeke a felismerés
- kriminalisztikai megközelítés
  - ☹ itt e beszélo érdekeivel ellentétes az, hogy ot felismerjék

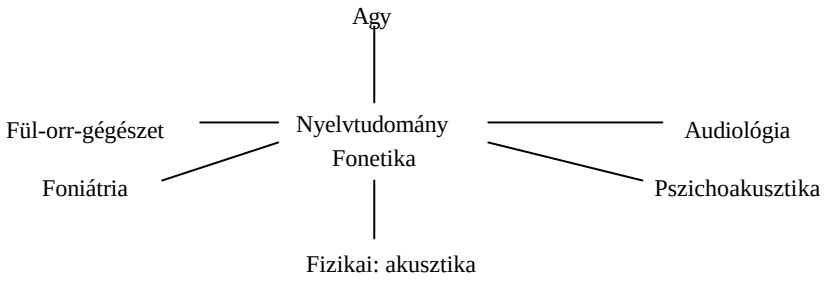
4.4.1.4. A beszélofelismerésnek alapvetoen két iránya van

- az n lehetoségbol kizárható-e nx személy
- az n lehetoségbol melyik az nx esemény
- a ketto kombinációja: benne van-e, és ha igen, ki lehet?

4.4.1.5. A beszélofelismerés paradoxona



4.4.1.6. A témához kapcsolódó tudományágak:



4.4.2. A fonetikán belül: hangszínezet (ez alapján azonosítja a beszélot)

- milyen mértékben jellemzo az emberre a hangja, beszéde (feltételezés: teljes mértékben)
- miképpen határozható meg az egyéni hangszínezet
- mely beszédképzési konfigurációval mutatja a legszorosabb kapcsolatot
- a zönge, a toldalécso, artikulációs mozgások vagy az összes együtt
- Miként fejezheto ki: akusztikai-fonetikai, percepciós-fonetikai vagy mind együtt
- mindennapi életben: szubjektív benyomások a hangról (minden embernek vannak bizonyos elvárásai, amik csak akkor tudatosulnak, ha valami nem stimmel, nem illik a környezetbe), a hallgatónak bizonyos feltételezései vannak a beszélorol

4.4.3. Beszédprodukció

- makrotervezés
- gondolati folyamat ↔ ismeretek, fogalmak
- pragmatika (hol mondom)
- szemantikai szerkezet, jelentés (milyen kifejezéseket válogatok ki) – stratégiák
- mikrotervezés
- szintaktikai szerkezet (milyen sorrendben, hogyan mondom oket) – transzformációs szabályok
- feszíni szerkezet – fonológiai szabályok
- fonetikai szerkezet
- artikulációs muködések
- akusztikai hullámforma

4.4.4. További problémák

- egyetlen beszélo egyetlen hangja laboratóriumi körülmények között és egyéb tényezok közül többet is elhanyagolva is nagyon sokféle lehet
- felpattanó zárhangok esetén a zöngésedési ido (VOT [ms]) nagyon jellemzo az adott mássalhangzóra (p, t, k), azonban a határértékek nagyon széles tartományokba esnek
- a spontán beszédben eltunik a szó, a magánhangzók nagy rész ’svá’, azaz ’ö’ lesz a kiejtés pongyolasága miatt (nem tökéletes hangképzés)
- az alaphang magasságának változása: a korral változik
  - a noké felnottkorig kicsit mélyül, de alapvetoen nem változik, idosebb korra jobban mélyül
  - a férfiaké kamaszkorban nagyon mélyül, felnottkorban egész mély, majd idosebb korban újra magasodik
- a beszéd akusztikájának relatív állandósága
  - egy nap elteltével fobb tényezok tekintetében a két kiejtés akusztikai jegyei között nincs különbség, de a részletekben már nagy eltérés mutatkozhat
  - gyermek és felnottkorban láthatóan eltéro
  - 20 év alatt az invariáns elemek nem változnak, de a hangszínkép között jelentos különbségek vannak
  - fiatal és idos között még nagyobb különbségek

- az érzelmek is befolyásolják (öröm, bánat, stb., de ez a ketto azonosítható a legjobban)
- a prozódia jellemzobb, mint a szegmentális szerkezet: alaphang-magasság, tempó, intenzitás, szünetstruktúra, ritmus, artikulációs változás
- alkoholos állapot: bizonyos szintig nincs hatása a beszédre, a mennyiség egyénfüggo
  - beszédtempó lassul
  - alaphang magasabb lesz
  - a szünetek elofordulása és idotartama no
  - a beszéd intenzitása no
  - az artikulációs mozgások elnagyolódnak
  - a beszéd tempójának változása
- a beszéd tempójának változása: meghatározó a város, környezet.
  - 1869 óta vannak erre adatok: 26-ról 1995-ig 65 szó / percre növekedett a beszéd sebessége
  - ami a beszédgyorsulást meggátolja: a hallgató megértése
  - ha automatizálni szeretnénk a beszélofelismerést, az idoviszonyok változása gondot jelenthet: két beszédminta tempója nem azonos, akkor most ugyanaz a beszélo volt-e vagy sem, továbbá befolyásolhatja a beszéd sebességét az érzelmi állapot és legfoképpen a zajviszonyok.

- környezeti zaj hatása a beszélore
  - beszédtempó gyorsulása
  - alaphangmagasság no
  - intenzitás no
  - monoton jellegu lesz a beszéd (moduláció csökken)
  - hangsúlyhibák gyakoribb elofordulása
  - ejtéshibák (szegmentális, szupraszegmentális), pongyola kiejtés, hanghibák
- tempó
- artikulációs tempó: mennyi hasznos beszédjel esik egy adott idotartamra (beszédképzésre fordított ido)
- beszédtempó: a nem hasznos beszédjeleket is beleértve (megakadás, szünet, 'ö', ismétlés, levegovétel)

Olyan esetek is elképzelhetok, amikor a teljes beszédbol csak 300 – 3500 Hz-re sávkorlátozott spektrum áll rendelkezésre (tipikusan telefon)

- redundancia csökken
- hiányoznak bizonyos invariánsok
- sokkal zajosabb a jel

4.4.5. Használt akusztikai vizsgálatok

- az [e] formánsszerkezeteinek különbözosége: alsóbb indexszámú formánsok között nincs különbség, de a magasabbakban igen, továbbá a formánsok sávszélessége is mutathat változásokat
- LPC analízis jó segítség lehet
- felhang szerkezete (telefon és studiofelvétel esetén, de a teljes spektrum alapján nem lehet eldönteni)

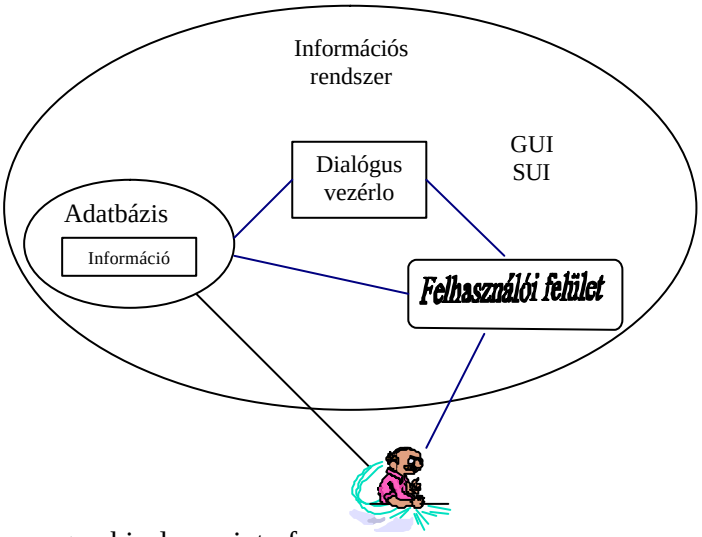
4.4.5.1. Akusztikai-fonetikai paraméterek a beszélo felismerésben

- központi formánsfrekvenciák és átmenetek
- maximumpontok
- formán sávszélessége
- rés és zárhangok zörejfrekvenciái
- sajátos spektrografikus alakzat
- felhangszerkezet
- magánhangzók idotartama
- beszéd és artikulációs sebesség
- csendes szünetek száma
- dallammenet
- hangsúlyozás
- egyéni ejtési sajátság akusztikai tükrözodése
- áthangolt spektrum (LPC analízis)
- Mi a helyzet a torzítással (amikor szándékosan el akarja változtatni a hangját)
  - az átlagember erre képtelen
- életkor, termet, súly meghatározása hang alapján (életkor elég jól meghatározható, termet kevésbé, de a súly nem jellemzo
- olyan beszédtorzítás, ami nem ismerheto fel (jelenleg): suttogás

5. BESZÉDINFORMÁCIÓS RENDSZEREK

- alapveto kérdés: mi lesz a rendszer célja, kinek készül a rendszer, ki fogja használni
  - a rendszert alapvetoen a felhasználónak készítjük
- a beszédinformációs rendszer elemi építokockákból épül fel
  - mit rakunk össze: milyen elemeket használunk, jó elemek-e ezek
  - hogyan rakjuk össze: az építoelemeket megfeleloen rakjuk-e össze

5.1. Beszédinformációs rendszer felépítése

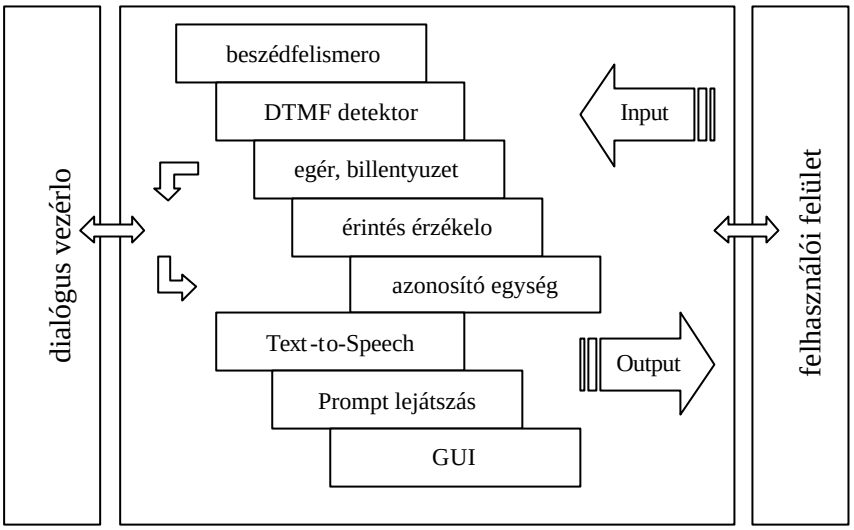


GUI – graphical user interface
SUI – speech user interface

- valamilyen információt valahogyan el kell juttatnunk a felhasználóhoz

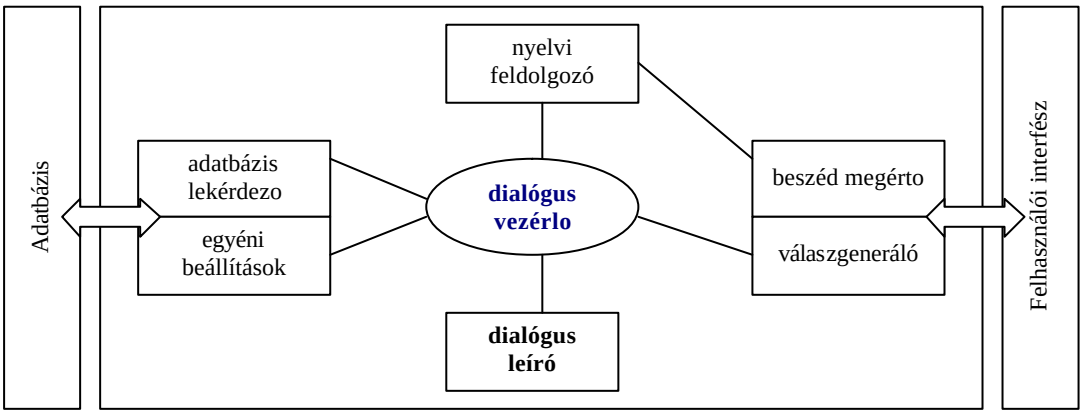
- a dialógusvezérlo arra szolgál, hogy a felhasználót rávezesse, hogy hogyan érje el ezt az információt
- a felhasználónak hozzá kell férnie az adatbázishoz, viszont a közvetlen adatbázis-hozzáférés leszukíti a felhasználók körét

5.1.1. A felhasználói felület



- a kimenetek beszédinformációs rendszer függoek: a felhasználó felol többféle kommunikáció lehetséges attól függoen, hogy milyen rendszert szeretnénk megvalósítani (beszéd, grafikai, egér, billentyuzet, érintésvezérlo
- az interaktív muködés érdekében: visszahallgatás
- prompt lejátszás: elore rögzített felvételek lejátszása, pl. „Üzenethallgatáshoz nyomja meg a 1. gombot!”
- az output lehet pl. bankjegykiadó, számlanyomtató

5.1.2. Dialógus vezérlo



- vezérlo
  - irányítja a feldolgozást (processzor, több gép, esetleg AI)
  - a beérkezo információ alapján a DB lekérdezot hívja meg
- válaszgeneráló
  - biztosítja a kapcsolatot: a kimeno egységek közül melyikkel és hogyan, mit kell kommunikálni
- felhasználói interfész felül érkezik valamilyem metanyelven az input (DTMF karakterek, beszédfelismerotol valamilyen adat)
  - a beszédmegérto: nem érvényes input kiszurése

- nyelvi feldolgozó: további egyszerusítések
- egyéni beállítások arra szolgálnak, hogy az egész dialógus egyéni lehessen a felhasználónak (rendszerfüggo illetve rendszerfüggetlen)
- dialógus leíró:
  - az inputok közül melyik micsoda, mikor mit várunk, mire mit kell csinálni (a protokoll)

5.1.2.1. A rendszer modalitása

- unimodális: 1 érzékre hat
- multimodális: több érzékre hat
- és ezeket input és output oldalról is vizsgáljuk
- PC
  - egér
  - billentyuzet
  - GUI
  - beszéd
- Információs pult
  - GUI
  - érinto képernyo
  - mutatás detekció
  - gesztus

5.1.2.2. Telefonos alkalmazás

- vezetékes (klasszikus: kizárólag hang továbbítása)
- mobil (régi, analóg)
- mobil (GSM, nemcsak hangátvitel, SMS, WAP)
- IP telefon
- Videotelefon

5.2. Dialógus rendszerek osztályozása

- A vezérlés jellege szerint (ki kezdeményez, ki irányítja a rendszer muködését)
  - rendszer vezérelt
  - felhasználó-vezérelt
  - vegyes kezdeményezésu
- a vezérlés módja szerint (elsosorban telefonos alkalmazásoknál)
  - DTMF
  - beszédfelismero
  - egyéb

5.2.1. Rendszer vs felhasználó vezérelt

- a rendszer határozza meg a navigációt
- menürendszer, felajánlott választási lehetoségekkel (pl. hangposta: mindig végig kell menni az összes menüponton)
- a felhasználó határozza meg a navigációt (pl. úgy döntök, hogy az utolsó két üzenetet törlöm)
- vegyes kezdeményezésu (lehetoség a navigáció módosítására, az elozo ketto kombinációja)

5.2.1.1. Menürendszer tervezési szempontok

- építokockák felhasználása (ha van egy muködo menü, abból már lehet építkezni)
- 4 választási lehetoség (ennél több ne legyen, mert a felhasználó úgysem tudja megjegyezni), visszalépés biztosítása
- maximális mélység 4-5 szint
- felhasználófüggo menürendszer (egyértelmu funkciók, következetes rendszer)
- csak a témakörhöz tartozó információk közlése
- megfelelo részletesség

- újdonságok, fontos elemek kiemelése
- gyakran használt gombok: kényelemhez legyenek igazítva
- egyes gombokat úgy helyezzünk el, hogy többletjelentést is lehessen nekik tulajdonítani (pl. nyilak)

5.2.1.2. Rendszerfüggetlen egyéni opciók

- felhasználói szint
  - kérdések, magyarázatok hossza, részletessége
  - választási lehetoségek száma
  - felajánlott választási lehetoségek száma
- felolvasás paraméterei
  - beszélo kiválasztása (férfi, no)
  - beszédtempó beállítása
  - szünetek idotartama (pl. mondatok között)
- adaptív változtatás / felhasználó vezérelt

5.2.1.3. DTMF vezérlés

- Dual Tone Multi-Frequency (4*4 frekvencia)
- adatbevitel a telefon billentyuzetén
- ☺ nagyon megbízható
- ☺ kipróbált technológia
- ☺ olcsó
- ☹ a kialakítható menü nem felhasználóbarát
- ☹ nehézkes a használat, ha a billentyuzet nem elérheto
- ☹ humán operátor szükséges lehet

5.2.1.4. Vezérlés hanggal

- ☺ a telefonba beszélni természetes dolog
- ☺ szélesebb témakörben alkalmazható (nemcsak számok)
- ☹ megbízhatatlanabb
- ☹ bizonyos esetekben jóval lassabb, mint a DTMF (foleg adatbevitel illetve gyakorlok felhasználók esetén)
- ☹ kötött szókészlet
- Yes/No elvu rendszerek
  - lassú és természetellenes
  - legmegbízhatóbb a beszédfelismeros rendszerekben
  - jól kell megválasztani a yes/no magyar megfelelojét
  - 2 szavas szótár nem elég (igen, jó, OK, mehet, rendben, ja, aha …)
  - adatbevitel fa struktúrában
- Kötött szótáras
  - meghatározott (10-20) szó felismerése
  - kényelmesebb menürendszer jellegu
  - beszélofüggetlen / adaptív jellegu (beszélofüggetlen mag és a felhasználóhoz igazítják)
  - gyors elérés mély struktúrák esetén is
  - keverheto a DTMF vezérléssel „fall back” (visszalépünk DTMF rendszerre)
  - bizonyos esetekben az adatbevitel nehézkes (számlaszám, visszaellenorzés, javítás)
- adott témakörben bármilyen információ mondható. lekérdezheto(a magyar sokkal szabadabb, jól meg kell gondolni, hogy mit próbálunk felismerni)
- felhasználó-vezérelt
- emberközeli használat
- pl. menetjegy-árusító rendszer
- diktáló rendszerek (tematika nélkül)
- mesterséges intelligencia szükséges (komoly nyelvtani szabályok, komplex dolgok)
- beszédfelismerés + mozgás és kézmozdulatok felismerése

5.2.1.5. Megerosítés (verifikáció)

- szükséges (mivel nem 100%-os a megbízhatóság, hibás bevitel)
- elvárt (a felhasználó biztonságérzete miatt): lehet explicit pl. közvetlen visszakérdezéssel és implicit pl. elrejtve a következo kérdésben
- explicit megerosítés minden egyes adatra: rákérdezés minden egyes adatra
  - eldöntendo kérdés
  - egyszeru struktúra
  - kényelmetlen a dialógus a felhasználó számára
- explicit megerosítés javítással: rákérdezés minden egyes adatra
  - ugyanaz, mint az elobb
  - igen/nem válasz mellett a javított adat is megadható
  - gyorsabb dialógusmenet
  - kevésbé akadozó

- explicit megerosítés több adatra: rákérdezés minden adatra egyszerre
  - kevesebb kérdés
  - igen/nem válasz mellett a javított adat is megadható / nem adható
  - természetesebb
  - csak az adatbevitel végén van ellenorzés
  - gondot jelenthet, hogy jól értelmezi-e a rendszer a mondanivalót
  - megnohet a kérdés összetételének bonyolultsága
- implicit megerosítés
  - következo adatbekérésbe ágyazott ellenorzés – a kérdésben benne van a régi információra vonatkozó visszakérdezés + az új információ
  - közelebb áll a természetes párbeszédhez
  - a kérdés hossza megno
  - javítás nehézkesebb
  - a rendszer bonyolultabb
- hibás megerosítés
  - felismerési probléma
  - rosszul ismeri fel
  - a felhasználó nem igen/nem-mel válaszolt
  - nem ismeri el a rendszer
  - többszörös megerosítés kritikus adatoknál: nem rögtön utána, inkább a végére érdemes berakni még egy megerosítést

5.2.2. Dialógus leíró eszközök

- SAPI
- VoCAPI
  - kisméretu eszközök: telefon, mosógép, fénymásoló
- ECTF – megpróbálja összefogni a különbözo technológiákat
- VoiceXML