A. Basics

1. Hang:

- Mechanikai rezgés, amely hullám formájában terjed a levegőben vagy más közegben, és az emberi fül érzékelni képes. Analóg, folyamatos hullám/jel.

A hang keletkezése:

- A nyugalmi állapotából kimozdított tárgy (rugalmas test) rezgésbe jön, és az ebből származó energiát átadja a környező levegőmolekuláknak, melyek szintén oszcillálni kezdenek. Ez az oszcilláció helyenként sűrűsödés-ritkulás nyomásváltozást hoz létre a közegben, ez a váltakozó légnyomás kis nyomáshullámot formában terjed tovább. Ezek a nyomáshullámok, melyeket hanghullámoknak nevezünk, a levegő (vagy más rugalmas közeg) részecskéi közötti gyors, periodikus nyomáscsökkenéseket jelentenek, amelyek a hang terjedésének alapját képezik.
- Amikor eléri a fülünket, a dobhártyát is rezgésbe hozza, amit az agyunk hangérzetként értelmez.

1. Mechanikai hullám:

- Olyan hullám amely oszcillál, és egy olyan rezgés amely terjedéséhez valamilyen anyagi közeg szükséges.

Típusai:

a) Longitudinális hullám: a részecskék rezgése párhuzamos a hullám haladási irányával. (pl. hang a levegőben)

b) Transzverzális hullám: a részecskék rezgése merőleges a hullám haladási irányával. (pl. húr rezgése, fény elektromágneses hullámként)

c) Felületi (vegyes) hullám: olyan hullámfelületeken terjed, ahol a részecskék mozgása körpályán vagy ovális pályán történik. (pl. víz felszínén a hullám)

2. Elektromágneses hullám:
- Nincs szükség anyagi közegre, vákuumban és terjed.
(pl. fény, rádióhullám, röntgensugárzás)

3. Egyéb hullámok: Anyaghullám, Gravitációs hullám

*vákuum - légüres tér, anyagban szegény tér

Közeg típusok

1. Homogén közeg:
- minden pontjában ugyanolyan (pl. desztillált víz, tiszta levegő)

2. Inhomogén közeg: különböző pontokban eltérő fizikai tulajdonságú
(pl. sós víz levegő, tengervíz stb.)

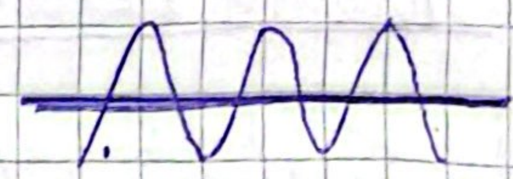
3. Anyagi közeg: minden fizikai test, mely tömegek és térfogattal bír. (szilárd, folyékony, légnemű)

4. Nem anyagi közeg: vákuum

11. Hullámforma: egy fizikai jelenség értékének időfüggő vagy más változóhoz viszonyított grafikus ábrázolása.
A jel fizikai alakját egy változását mutatja meg.

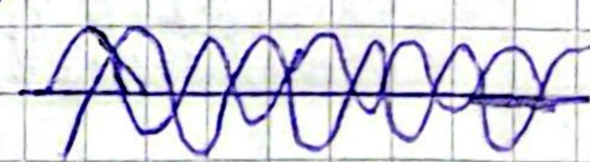
1. Periodikus: ismétlődő mintázatot mutatnak

- Egyszerű - Egyetlen szinuszhullám:



- tiszta szinuszos jel, amelynek csak egy meghatározott frekvenciája van, harmonikus felhangok nélkül. (pl. tiszta hangvilla)

- Összetett - Több szinuszhullám

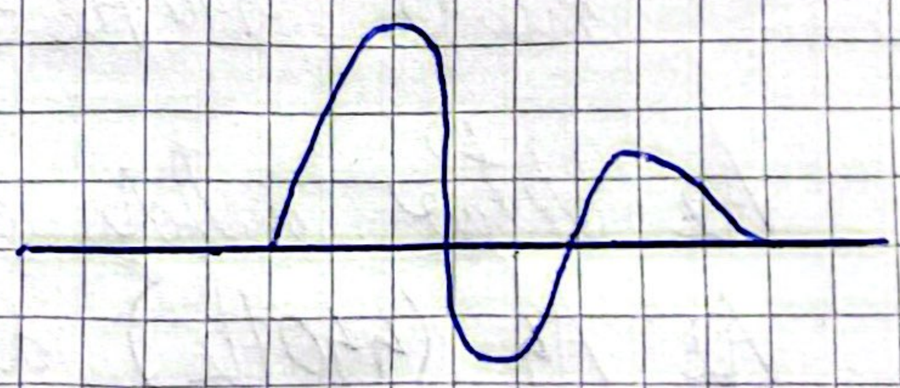


- több különböző frekvenciájú szinuszhullám összegéből alakulnak ki. (pl. hangok/zenei hangszerek, melyek hangzása sokféle frekvencia jelenlétéből adódik.)
- Az ilyen hullámokat „Fourier-analízis"el elemzik.

2. Aperiodikus: nem mutatnak rendszeres, ismétlődő mintázatot

- Folyamatos - Zaj: teljesen rendezetlenek, különféle típusú zajként (barna, fehér, rózsaszín zaj) jelennek meg. (pl. rádió statikus zaja, bizonyos elektronikai eszközök működése) Nem lehet periodikus jellegű bontani, főként véletlen eseményekből tevődnek össze.

- Tranziens - Pulzus: rövid, menn ismétlődő impulzusok, hirtelen keletkezhetnek és gyorsan lecsengenek. (pl dobütés, villám, kopogás, EKG)



Hallástartomány: Az emberi hallástartomány kb. 20Hz-20000Hz

20 Hz alatt infrahang
20000 Hz felett ultrahang

Fizikai hullámjellemzők:

- Frekvencia (ν/f/1/T): egy esemény ismétlődése egységnyi idő alatt
- Amplitudó (A): a legnagyobb kitérés az egyensúlyi vagy nyugalmi helyzettől számítva
- Periódus (T): az időtartam amely alatt egy periodikus jelenség teljes ciklust végrehajt.
- Fázis (φ): a hullám aktuális állapota vagy kezdőpontja az időben vagy más hullámokhoz képest.

Hangérzékelési jellemzők: hangerősség, hangmagasság, időtartam, színezet

Az emberi hallás logaritmikus, nem lineáris. Arányosnak érzékeny, ha két hang frekvenciájának aránya azonos, az egymásnak hangzónak hallatszik, függetlenül az abszolút értékektől.

pl.: 110 Hz - 220 Hz egymásnak hangzós
mint 440 Hz - 880 Hz

Az okta miatt egy hangfrekvencia duplázódik
Az A4 (440Hz) a nemzetközileg elfogadott referencia
Szinte minden hangszer, zenekar és digitális zenei eszköz (MIDI) ehhez igazodik, és ettől számítva a többi hangot.

$$F(p) = 440 \cdot 2^{\frac{p-69}{12}}$$

A fenti képletet alkalmazzák egy adott MIDI pitch hangszámhoz tartozó pontos hangfrekvencia meghatározásához. Bármilyen hangmagasság pontos akusztikai frekvenciája kiszámítható - az összes digitális zenei rendszer (MIDI, DAW, VST) ezt alkalmazza.

$$F(p+1) / F(p) = 2^{\frac{1}{1200}} = 1{,}059 \Rightarrow 100 \text{ cent}$$

- két egymás utáni MIDI hang frekvenciának aránya

<u>Cent</u> egy nagyon apró léptékű egység a hangmagasság mérésére, amelyek a hangközök nagyon pontosan kifejezhetők.

egy oktáv 1200 cent
egy félhang 100 cent

Emberi számára hallható különbség kb. 10-25 cent

pl. ekkor $f_2 = f_1 \cdot 2^{\frac{\text{cent}}{1200}}$

$$f_2 = 440 \cdot 2^{\frac{1}{1200}} \approx 440.255 \text{ Hz} \Rightarrow +1 \text{ cent}$$

III. Hangteljesítmény (W): a hangforrás mennyi energiát bocsát ki minden irányba, időegységenként.

Hangintenzitás (W/m^2): egy adott felületre mennyi hangenergia jut egy másodperc alatt.

Hallásküszöb (Threshold of hearing):

$$TOH = 10^{-12}\ W/m^2 \quad (=) \quad 0 \text{ decibel}$$

- a hang minimális intenzitása amit az emberi fül még érzékelni képes, a hallásérzékelés kezdőpontja. Innen számoljuk a decibelt

Fájdalomküszöb (Treshold of pain):

$$TOP = 10 \cdot W/m^2$$

Intenzitásszint (dB): megmutatja egy adott hang mennyiszer erősebb a lehetőleg még hallható hangnál képest (TOH), ahol a viszonyítási alap a hallásküszöb. Logaritmikus skálán, decibelben. Arány két hangintenzitás közt.

$$dB(I) = 10 \cdot \log_{10}\left(\frac{I}{I_{TOH}}\right)$$

I - vizsgált hangintenzitás $\qquad I_{TOH}$ - hallásküszöb

1. hallásküszöb intenzitása

$$I = I_{TOH} = 10^{-12}\ W/m^2$$

$$dB = 10 \cdot \log_{10}\left(\frac{10^{-12}}{10^{-12}}\right) = 10 \cdot \log_{10}(1) = 0\ dB$$

- a hallásküszöbön a hangintenzitás szintje 0 dB

2. normál beszédintenzitás 1kHz

$I = 10^{-6}\ W/m^2$

$dB = 10 \cdot \log_{10}\left(\frac{10^{-6}}{10^{-12}}\right) = 10 \cdot \log_{10}(10^6) = 10 \cdot 6 = 60\ dB$

3. hangos koncertterem

$I = 10^{-3}\ W/m^2$

$dB = 10 \cdot \log_{10}\left(\frac{10^{-3}}{10^{-12}}\right) = 10 \cdot 9 = 90\ dB$

4. fájdalomküszöb körül

$I = 1\ W/m^2$

$dB = 10 \cdot \log_{10}\left(\frac{1}{10^{-12}}\right) = 120\ dB$

Nagyjából 3 dB után az intenzitás megduplázódik

Hangosság

- A hangintenzitás szubjektív érzékelése, befolyásolja az életkor, az egyén, a hang frekvenciája és időtartama. A közepfrekvenciás hangok hangosabbnak érződnek.

Mértékegysége a phons.

Hangszín: olyan tulajdonság amely megkülönböztet két azonos hangmagasságú és hangerejű hangot egymástól. Főként az határozza meg, hogy a hangforrás milyen felharmonikusokat és zajokat bocsát ki, milyen arányban az alapfrekvencián kívül. - Többdimenziós: többféle módon és alkalmasságban tényező alkotják

- Hangfelfutás: elindulás (attack), elhalás (decay), kitartott (sustain), elhalás (release)
- Felharmonikusok: más-más arányú felharmonikusok az alapfrekvencián kívül
- Amplitudó és frekvenciamoduláció: erősség-magasság hullámzása időben

Komplex hang: nem egyetlen tiszta, szinuszos rezgésből, hanem több különböző frekvenciájú rezgés összege. Ezek a különálló rezgések a hang alapfrekvenciájától és annak egész számú többszöröseiből tevődnek össze.

1. alapfrekvencia határozza meg a hangmagasságot

pl: $f_1 = A_4 = 440\ Hz$

2. hellamonitusok az alapfrekvencia egész számú többszörösei

$f_2 = 880\ Hz$ $f_3 = 1320\ Hz$ .... stb.

Ezek a hellamonitusok határozzák meg a hangszínt (hellamonikus hangszerek esetén pl: hegedű, ének, trombita)

Inharmonikus hangszerek esetén a rezonanciafrekvenciák nem egész számú többszörösei az alaphangnak
pl: harang, csiptető, triangulum, dob, xilofon

Hangkeltés (gerjesztés) szerinti osztályozás

Idiofonok - hangszertest rezeg (harang, xilofon)
Membranofonok - kifeszített membrán rezeg (dob, kanna)
Kordofonok - kifeszített húr rezeg (gitár, zongora, hegedű)
Aerofonok - levegőoszlop rezgése (furulya, orgona)
Elektrofonok - elektronikus úton (szintetizátor)

B. Audio Signals for Machine Learning

1. Audio jel: elektromos jel amely lehetővé teszi (analóg vagy digitális) átvitelét hangfrekvencián. Tartalmazza a hang legfontosabb információit, mely lehetővé teszi a rögzítést, továbbítást és reprodukcióját.

Analóg jel:

- Folytonos időben és értékben, elméletben végtelen számú értéket vehet fel.
- Fizikai világ szinte minden jele analóg módon érhető
- Zajérzékeny, bármilyen kis torzulás befolyásolja a jel minőségét

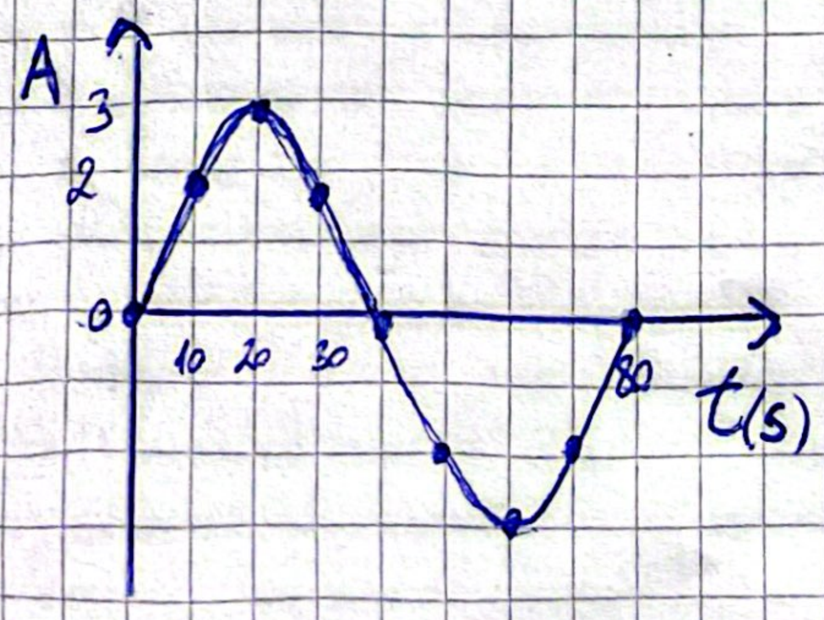
Digitális jel:

- Meghatározott pillanatban, meghatározott értéket vehet fel.
- Könnyű kezelés, hibák és zajok kevésbé befolyásolják
- A diszkrétizáció (kvantálás) miatt részleteket vesztőhet, véges számú

Analóg - Digitális átalakítás (ADC)

Pulse code modulation (PCM): olyan digitális módszer mellyel folyamatos jeleket alakítunk digitális formává (ADC)

1. Mintavételezés (Sampling): analóg jelet rendszeres periódusonként mintázzuk és az adott pillanatban mért amplitudót rögzítjük.



$t_n = n \cdot T$ $\quad t_5 = 5 \cdot 10 = 50\ s$

$f = \frac{1}{T} = 0{,}0125\ Hz$

Nyquist frekvencia az analóg jel legmagasabb frekvenciájának legfeljebb a mintavételi frekvencia felének kell lenni.

$$f_s = \frac{1}{dt} \qquad f = \frac{1}{T} \qquad f_N = \frac{f_s}{2}$$

$$f \leq f_N$$

előző példa esetén:

$$f_s = \frac{1}{10} = 0{,}1\ Hz \qquad f = \frac{1}{80} = 0{,}0125\ Hz \qquad f_N = 0{,}05\ Hz$$

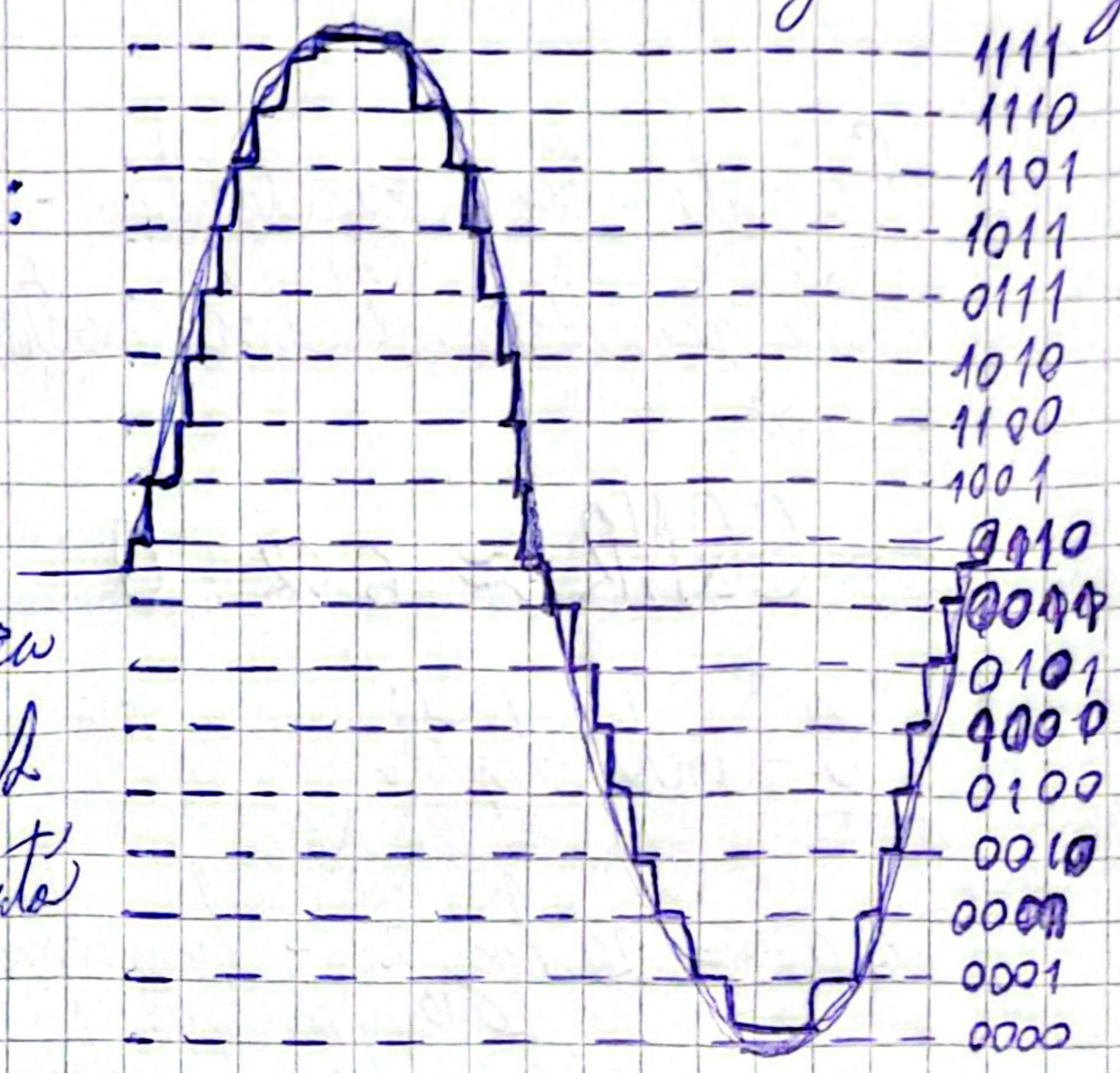
$$f \leq f_N \qquad 0{,}0125 \leq 0{,}05 \text{ megfelel}$$

- CD minőségű hangfelvétel esetén a mintavételi frekvencia 44100 Hz, mert ebből 22050 Hz a Nyquist frekvencia ami az emberi fül számára legmagasabb érzékelhető frekvencia.

Aliasing: alulmintavételezés, mintavételi pontokban levett értékek alapján nem lehet egyértelműen megfontolni az eredeti jel frekvenciáját.

2 Kvantálás (Quantization) a folytonos (analóg) jelekre, amely elméletben tetszőleges értéket felvehet az amplitúdó tengelyén, csak bizonyos, előre meghatározott értékszintekből engedünk meg értékeket.

Ezek a kvantálási szintek:



Resolution = num of bits:
A digitális jel felbontását az alkalmazott bitek száma határozza meg. Minél több bitet használunk annál részletesebben meghatározható az analóg jel amplitúdója

Bit depth: megadja hány kvantálási szintre lehet felosztani az analóg jelet. (Fenti ábrán 4 bit = 16 szint)

pl: CD resolution = 16 bit = 65536 szint
=> magas minőségű és alacsony kvantálási zaj

Digitális memória 1 percnyi hangy tárolásához

Sampling rate = 44100 Hz (CD)
- minden másodpercben 44100 mintát veszünk a hanghullámból

Bit depth = 16 bit / channel — channel: mono / stereo
- minden egyes hangminta 16 bites felbontásban kerül eltárolásra.

$$(((16 \cdot 44100) / 1\,048\,576) / 8) \cdot 60 \cdot \text{channel} = 5.49\ \text{MB}$$

másodpercenként eltárolt bitek száma; osztás 1 megabitre; megabit -> megabájt; sec

Ezen a méreten bizonyos fájltípus pl: MP3, veszteségesen tömörítenek és eltárolják azon anyagokat amelyeket az emberi fül nem érzékel.

Dynamic range: megadja egy mérő- vagy jelrögzítő rendszer (mikrofon, digitális hangkártya, erősítő) mellett legnagyobb és legkisebb jelek közötti torzítás nélkül feldolgozni.

Signal-to-quantization-noise ratio: mennyire erős a hasznos jel a kvantálásból származó zajhoz képest

$$SQNR \approx 6.02 \cdot Q \Rightarrow 6{,}02 \cdot 4 = 24{,}08 \text{ dB}$$

Q - bit depth

6,02 = 20 log 2 - logaritmikus skálán mért jel/zaj arány növekedése bitenként.

⇒ A jel kb. 24 dB-vel erősebb vagy tisztább mint a kvantálási zaj.

Hogyan rögzítjük a hangot?

1. **Hangjelek rögzítése**: a mikrofon az analóg hangrezgéseket elektromos jellé alakítja

2. **Analóg jel digitalizálása (ADC)**: a mikrofontól érkező folyamatos elektromos jelet egy digitális hangrögzítő eszköz diszkrét mintákra bontja, adott időközönként megméri

3. **Hangfájlok mentése**: digitális minták sorozatát hangfájlként tartja tárolásra. (WAV, MP3)

Hogyan reprodukáljuk a hangot?

1. Digitális jel fogadása: a digitális jel bekerül a DAC-ba egy elektromos jelet hoz létre, amely az eredeti analóg jelhez közelítő értékű.

2. Szűrés és simítás: az így kapott jel kezdetben lépcsőzetes alakú a diszkrétmintavétel miatt, ezért egy analóg aluláteresztő szűrő simítja a jelet, eltávolítva a magasabb frekvenciás zajokat, hogy folytonos, sima hullámformát kapjunk.

3. Kimenet a hangszóróhoz: a simított, analóg jel továbbjut a hangszóró erősítőjébe majd magára a hangszóróba, amely azt mechanikai hullámokká, azaz hanghullámokká alakítja.

II. Audio jellemző típusai

1. Absztrakciós jellemzők:

a. Low-level: legalapvetőbb matematikai-technikai jellemzők energy, spectral flux, spectral centroid, zero-crossing rate, amplitude envelope

b. Mid-level: összetett technikai szintű jellemzők fluctuation pattern, pitch- and beat-related descriptor, MFCC

c. High-level: összetett információk, zenei-esztétikai élmény key, chord, melody, rhythm, tempo, lyrics, genre, mood

2. Időbeli lefedettség: audio jellemző milyen időtartamot fed le

a. Pillanatnyi/Azonnali (~50 ms):

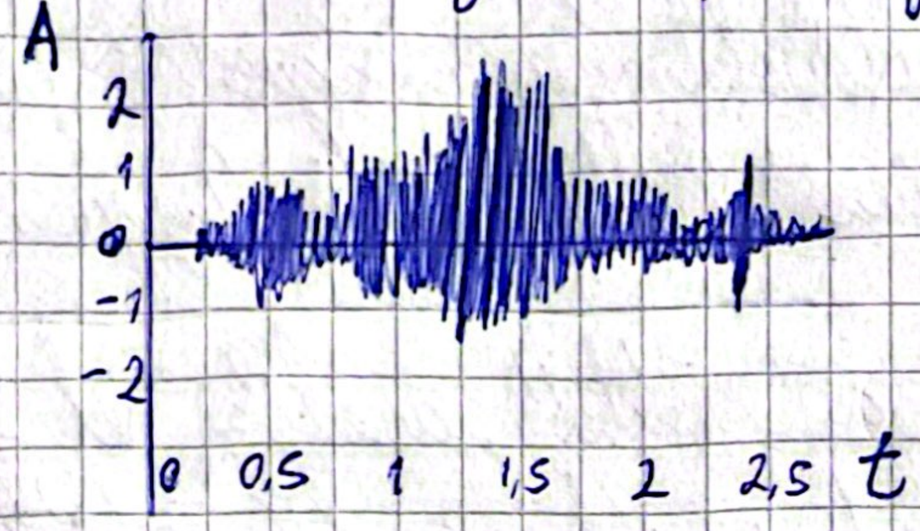
b. Szakaszos szint: néhány másodperc

c. Globális: teljes mű vagy hangjegyzet

3. Zenei szempont: esztétikai elemei egy adott hangnak ütem, hangszín, hangmagasság, harmónia, időtartam

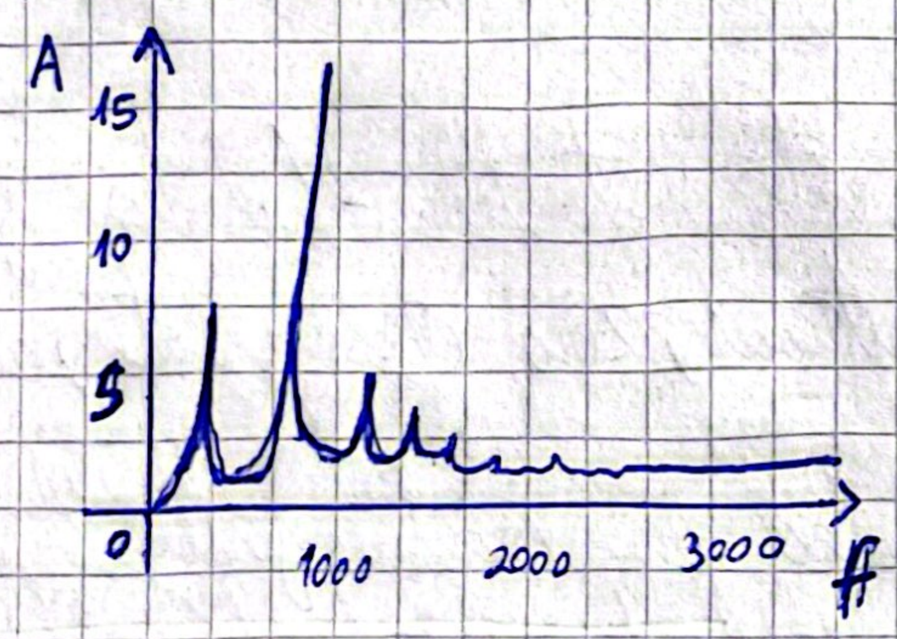
4. Szignál domain

a. Time domain: a hangot úgy vizsgáljuk, ahogy időben változik. A jel amplitúdója az idő függvényében változik



Hullámforma amely mutatja milyen hosszú vagy hangosabb a jel.

b. Frequency domain: a jelet frekvenciáira bontva (pl. Fourier-analízissel vagy spektrumával) vizsgáljuk. Ennek lényege, hogy megvizsgálja magasságon hangszínek vagy frekvenciákból tevődik össze a jel.



A jel mely frekvenciákat tartalmaz egyenletes vagy eltérő arányban és mely frekvenciakomponensek vannak jelen

6. Time-frequency representation: általában spektrogramon ábrázolja megmutatja egy jel milyen frekvenciákat tartalmaz és ezek hogyan változnak időben. pl. akusztikai felismerésben

5. Gépi tanulás fogalomköre (ML):

a. Hagyományos gépi tanulás: megpróbálja az összes hangjellemzőt az idő- és frekvenciatartományban. Ezekből az előre definiált, jól kezelhetőnek szánt jellemzőkből készít egy jellemző-vektort, amelyet aztán valamilyen ~~döntési fával~~ tanító algoritmusokat arra lehetséges minta SVM, k-nn, döntési fák.

b. Deep learning - automatikusan megveszi és a jellemzőket és tanulja meg azokat. Nem kell előre definiálni, hogy mely jellemzőt fontosak, hanem a hálózat magától tanulja meg a bemenő adatokból.

Extracting Audio Features

1. Framing - keretezés: ahhoz, hogy a hang feldolgozható legyen gépi tanulással vagy jelfeldolgozással feldarabolják rövid időablakokra. A frame egy olyan kis hangrészlet, amely rövid, de elég hosszú ahhoz, hogy valamilyen zenei, beszédbeli vagy technikai jellemzőt ki lehessen számolni belőle.

pl. 1 sample @44,1 kHz = 0,0227 ms

duration 1 sample << Ear's time resolution (10 ms)

- Egy frame hossza, 2 hatvány számú minta: 256, 512, 102...
  tipikus frame hossz: 256 - 8192 db 6 ms - 185 ms
  x db egymás utáni érték vizsgálata 1 frameb...

$$d_f = \frac{1}{f_s} \cdot K \qquad d_f = \frac{1024}{44100} \approx 0{,}0232\,s \approx 23\,ms$$

d_f - egy frame időtartama

f_s - mintavételi frekvencia

K - frame hossza mintában

2. Feature computation: minden egyes frameből kiszámítjuk a fontos jellemzőket (energy, amplitude envelope, zero-crossing rate). Ezek olyan mérőszámok melyek leírják a hang részleteit és gépileg feldolgozhatók.

3. Aggregation a számolt jellemzők statisztikáit (átlag, median, GMM) képezzük a teljes anyagra. Egy tanított, jól működő rendszer kezelhető jellemző-érték, vektor vagy mátrix.

4. Feature value/vector/matrix Az összes számolt jellemzőből egy adatmátrixot kapunk gépi tanuláshoz, hangfelismerésre, elemzésre.

Fourier-transformation: matematikai eljárás, amellyel az időben jel összetett jellegét bonthatjuk fel egyszerűbb, periodikus összetevőkre, sin és cos hullámokra.

időtartomány → frekvenciatartomány

Mivel a digitális hangmintát diszkrétek, Discrete Fourier-T (DFT) vagy annak hatékonyabb változatát a Fast-Fourier-T (FFT) használjuk. Hasznos hangszín elemzésben, zajszűrésben, hang detektálásban és digitális jelfeldolgozásban.

Spectral leakage: Fourier transzformáció során fellépő jelenség. Amikor a vizsgált jel ~~[illegible]~~ nem tökéletesen periodikus az elemzés időablakon belül, az adott frekvencia beszivárog a szomszédos frekvenciákba, megnehezítve a frekvenciák pontos elválasztását és torzítva a spektrumot.

Windowing: ablakfüggvények használatával (pl: Hann, Hamming, Blackman) csökkenthető a szivárgás. Simítják az időablak szélén lévő éleket, csökkentve a diszkontinuitást.

Hann window: segít tisztább, kevesebb torzítást tartalmazó frekvenciaelemzést készíteni az időbeli jelből oszáltal, hogy a vizsgált jel időablakának széleit kiegyenlíti.

Előny: egyszerű, könnyen implementálható
jó egyensúly a spektrális felbontás és oldalfüggvény k[illegible]

Korlát: nem szünteti meg teljesen a spektrális szivárgást más ablakok pl: Blackman kevésbbé zajosítják, de komplexebbek.

Formula: $w(k) = 0{,}5 \cdot \left(1 - \cos\left(\frac{2\pi k}{K-1}\right)\right)$, $k = 1 \dots K$

$w(k)$ - ablak súlytényezője az adott mintánál (Hann ablak)

K - ablak hossza (minták száma)

k - aktuális minta indexe az ablakban

Windowed Samples: $S_w(k) = S(k) \cdot w(k)$

$S(k)$ - eredeti komplex amplitúdója

$w(k)$ - Hann ablak értéke

$S_w(k)$ - ablakkal súlyozott minták (Windowed Sample

Overlapping (átfedés): a windowing miatt a kiválasztott frame jelének értékei az ablak szélén nullára csökkennek. Ezt javítja az overlap, így a szomszédos frameket különálló blokkok helyett fedik egymást részben.

Pl.: 50%-os overlap azt eredményezi, hogy a következő frame az előző felével van átfedésben. Az biztosítja, hogy az ablakolás által csillapított szélső értékek más frameken újra megjelenjenek, javítva a jel időbeli folytonosságát és zökkenőmentes elemzését.

Time-domain feature pipeline

analóg audio jel → PCM → Framing →
→ feature computation → aggregation → feature vector

felhasználási terület: zaj, beszéd, egyszerű jellemzők

Frequency-domain feature pipeline

analóg audio jel → PCM → Framing → Windowing →
→ Overlapping → Fourier-transformation

felhasználási terület: zene, hangszín, komplex hangok