

Extracția de trăsături acustice

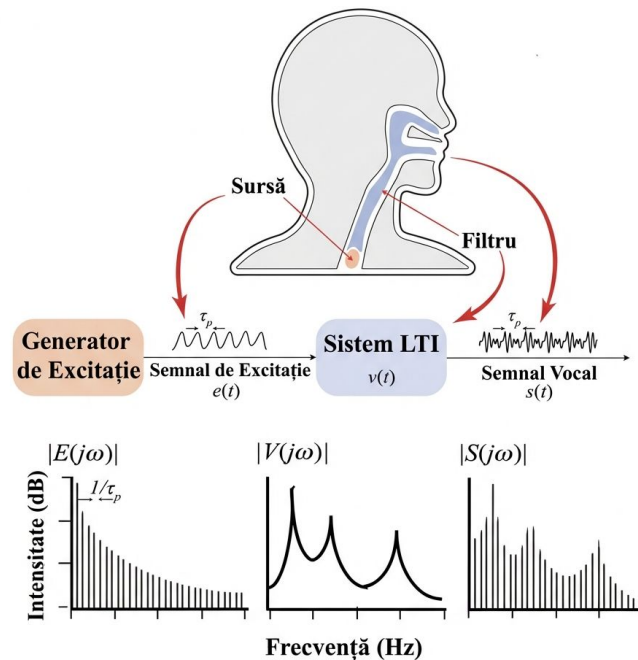
Prezentat de Tudor Bumbu

Cum producem vorbirea?

Pentru a ști cum să procesăm vocea digital, folosim **Modelul Sursă-Filtru**, care explică mecanismul fizic de producere a vorbirii.

Sursa (excitația) este originea sunetului brut. Poate fi vibrația corzilor vocale (pentru vocale) sau turbulența aerului (unele consoane: s, f, ș).

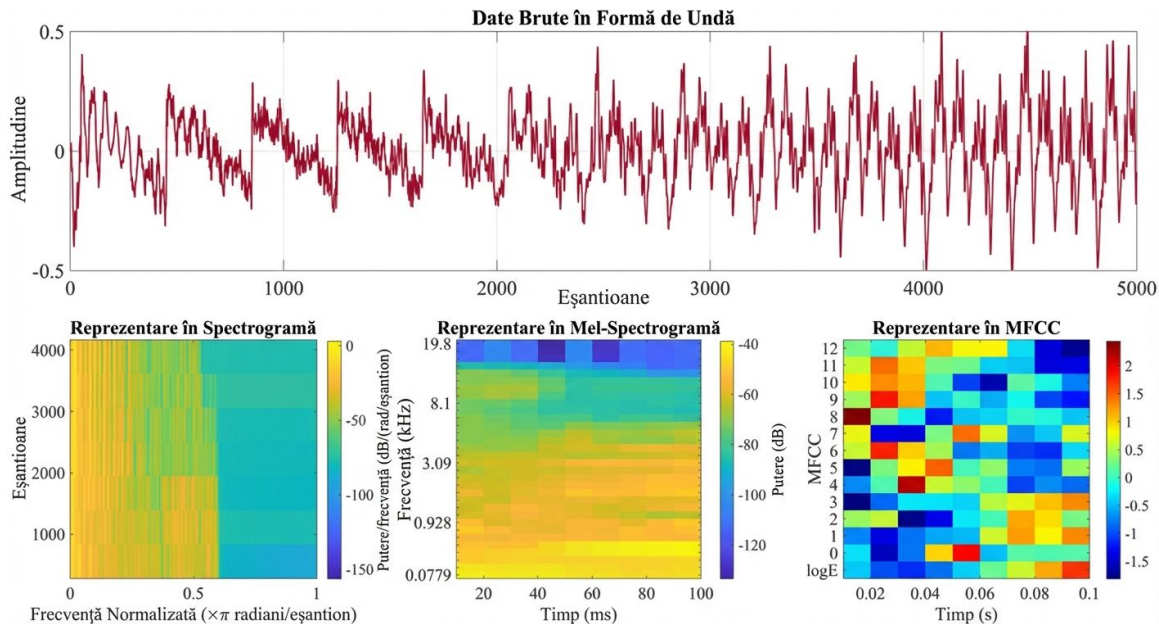
Filtrul (tractul vocal) include cavitatea bucală, poziția limbii, a maxilarului și a buzelor. Acesta „sculptează” sunetul brut.



Extracția de trăsături acustice

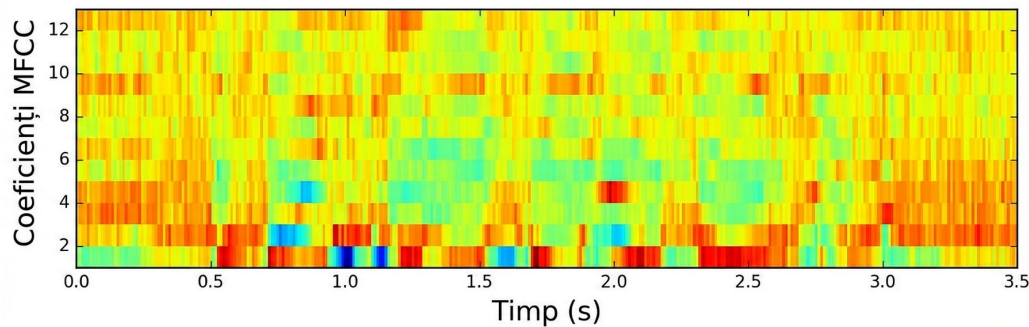
Prin *Extracția de trăsături* (feature extraction) transformăm semnalul într-o reprezentare:

- Compactă (mai puțină informație inutilă),
- Discriminativă (ajută modelul să distingă litere care sună asemănător).
- Robustă (stabilă indiferent de condițiile de înregistrare).



MFCC: Trăsături inspirate de auzul uman

Coeficienții Cepstrali pe Scara Mel (MFCC) reprezintă un standard în procesarea vorbirii care oferă o reprezentare compactă inspirată direct de percepția auditivă umană.

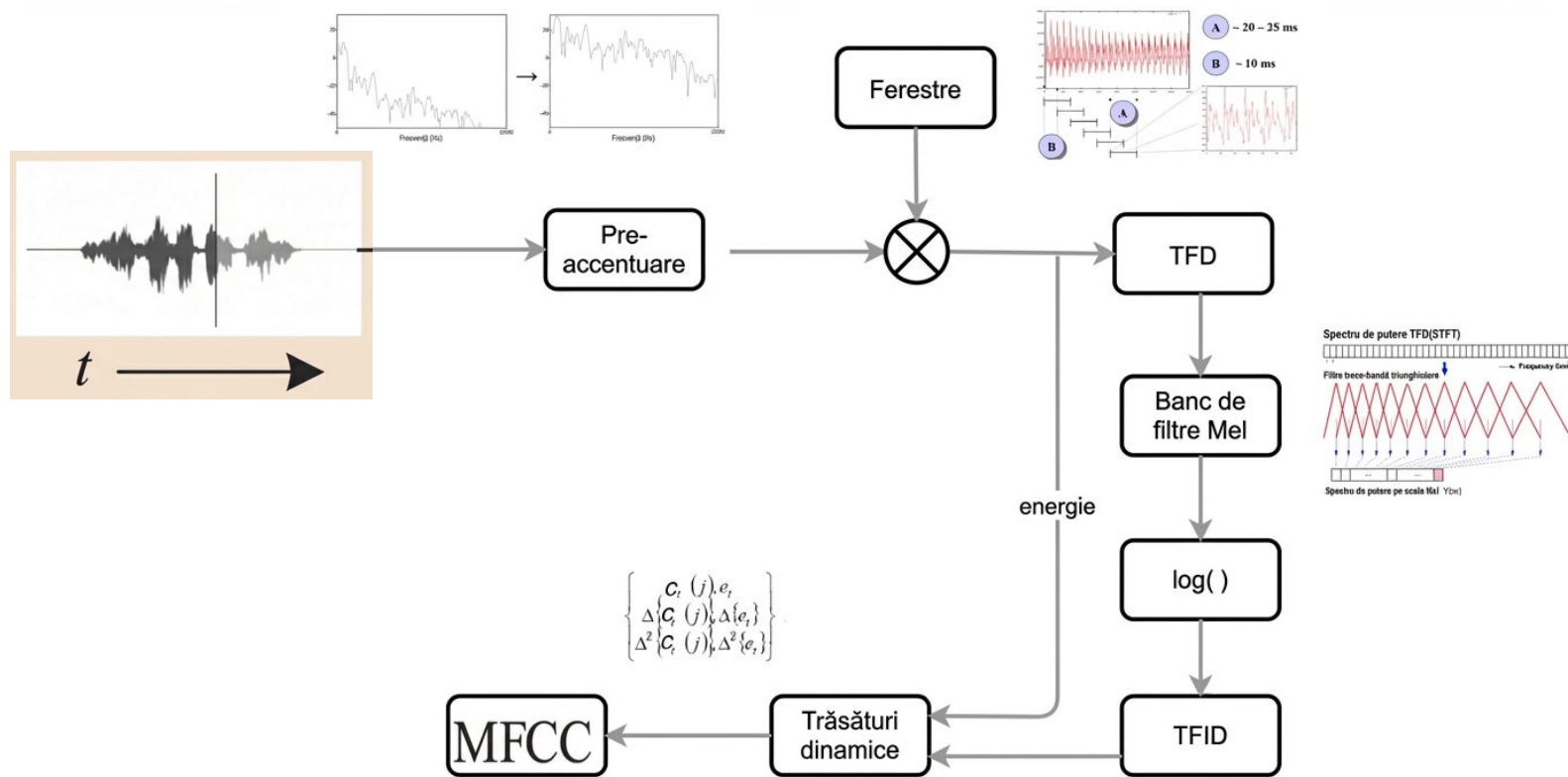


De ce folosim Scara Mel?

- La frecvențe joase, distingem foarte bine diferențe mici de sunet.
- La frecvențe înalte, avem nevoie de diferențe mult mai depărtate pentru a sesiza o schimbare.

Scara Mel modelează acest fenomen, cu filtre dese la frecvențe joase și mai late la frecvențe înalte.

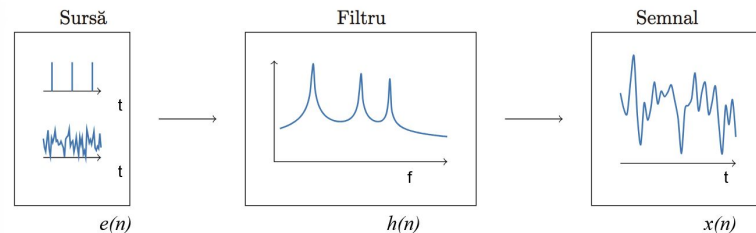
Cum se calculează MFCC?



LPC: Codarea Liniar Predictivă

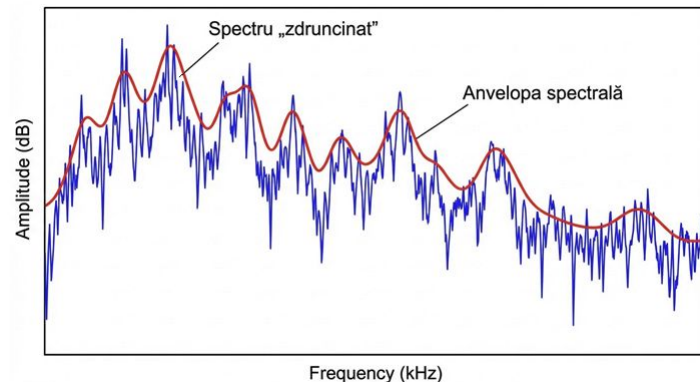
Spre deosebire de MFCC, metoda LPC este motivată de Modelul Sursă–Filtru.

- Prezice eșantionul următor din eșantioanele anterioare.
- Diferența = reziduul (sursa).
- Coeficienții LPC modelează anvelopa spectrală.



$$x(n) = h(n) * e(n)$$

Linia albastră = **spectrul real al semnalului vocal**
Linia roșie = **anvelopa spectrală**



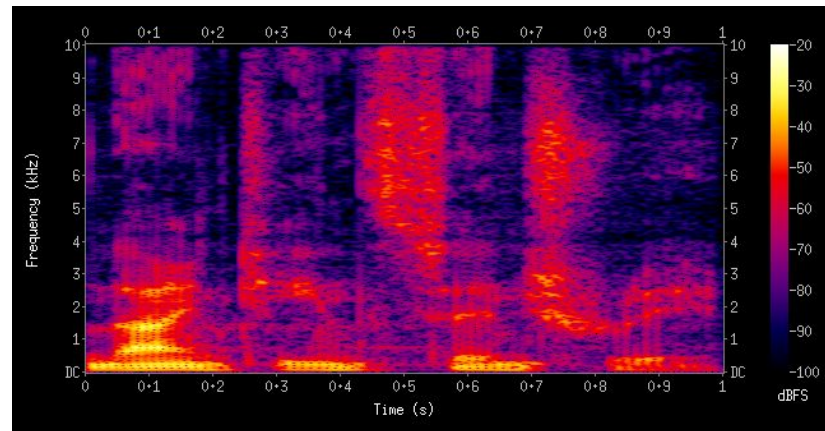
Spectrograma: Imaginea sunetului

Spectrograma este o reprezentare 2D a sunetului:

- X - timp
- Y - frecvență
- culoare - energie

Se obține prin STFT

(Transformata Fourier pe ferestre scurte succesive).



Tipare uzuale:

- Vocale - benzi orizontale intense (formanți)
- Consoane surde - zgomot difuz la frecvențe înalte
- Pauze - zone întunecate

Normalizarea datelor

Normalizarea mediei și varianței cepstrale (CMVN) este o tehnică de normalizare eficientă din punct de vedere computațional pentru recunoașterea vorbirii.

Pentru un coeficient cepstral dat $c_t(d)$, la cadrul temporal t și dimensiunea d , coeficientul normalizat $\hat{c}_t(d)$ se calculează astfel:

$$\hat{c}_t(d) = \frac{c_t(d) - \mu_d}{\sigma_d}$$