Preprocesarea Semnalului Vocal și Extracția de Trăsături

Acest capitol introduce fundamentele analizei semnalului vocal, punând accent pe modelarea teoretică a producției vorbirii, pe metodele de extragere a trăsăturilor acustice și pe tehnicile de normalizare destinate îmbunătățirii robusteții sistemelor. Vom începe cu o prezentare a modelului sursă-filtru, un cadru conceptual care a devenit piatra de temelie în procesarea vorbirii. Ulterior, vom detalia două dintre cele mai utilizate metode de reprezentare a semnalului: coeficienții cepstrali pe scară Mel (MFCC) și codarea liniar predictivă (LPC). În final, vom discuta rolul normalizării cepstrale și modul în care aceasta poate atenua variațiile nedorite introduse de canale sau de vorbitori diferiți.

2.1. Modelul Matematic al Producției Vorbirii: Sursă-Filtru

Modelul sursă-filtru, propus inițial de Gunnar Fant în lucrarea sa seminală "Acoustic Theory of Speech Production" (1960), este o abstractizare matematică a procesului de fonație umană care a devenit fundamentală în procesarea digitală a vorbirii. Acest model postulează că semnalul vocal, $s(t)$, poate fi descompus în două componente independente: o sursă de excitație, $e(t)$, și un filtru liniar, $h(t)$, care reprezintă tractul vocal (Flohberger, 2003).

Matematic, relația dintre aceste componente este exprimată în domeniul timp printr-o operație de convoluție:

$$s(t) = e(t) * h(t)$$

unde * denotă convoluția. Conform proprietăților transformatei Fourier, această relație devine o simplă înmulțire în domeniul frecvență:

$$S(\omega) = E(\omega) \cdot H(\omega)$$

unde $S(\omega)$, $E(\omega)$ și $H(\omega)$ sunt transformatele Fourier ale semnalelor $s(t)$, $e(t)$ și, respectiv, $h(t)$ (Titze, 2008).

2.1.1. Caracteristicile Sursei, E(ω):

Sursa modelează semnalul generat la nivelul glotei, conform teoriei vorbirii (Tokuda 2021).

Pentru sunete sonore (voiced): Sursa este un flux de aer quasi-periodic, generat de vibrația corzilor vocale. Acest semnal este modelat ca un tren de impulsuri. Spectrul său, $E(\omega)$, este format dintr-o frecvență fundamentală (F_0), percepută ca ton (pitch), și o serie de armonice (multipli întregi ai F_0) a căror amplitudine scade odată cu creșterea frecvenței.

Pentru sunete surde (unvoiced): Sursa este un zgomot aperiodic, generat de fluxul de aer turbulent care trece printr-o constricție în tractul vocal (de exemplu, pentru sunetele /s/ sau /f/). Acest semnal este modelat ca un zgomot alb, iar spectrul său, $E(\omega)$, este aproximativ plat pe întreaga bandă de frecvență.

2.1.2. Caracteristicile Filtrului, H(ω):

Filtrul reprezintă funcția de transfer a tractului vocal, care acționează ca un rezonator acustic. Forma sa este determinată de poziția articulatorilor (limbă, maxilar, buze). Funcția de transfer H(ω) nu este plată; ea are vârfuri la anumite frecvențe, numite formanți, care corespund frecvențelor de

rezonanță naturale ale cavității vocale. Acești formanți amplifică armonicile sursei care se află în apropierea lor și le atenuează pe celelalte, modelând astfel anvelopa spectrală a sunetului final și determinând identitatea fonetică a vocalelor.

Ipoteza fundamentală a acestui model este independența sursei de filtru, ceea ce permite analiza separată a acestora. Deși studiile moderne arată că pot exista interacțiuni non-liniare între sursă și filtru, în special în producția vocală a cântăreților, această ipoteză de independență este o aproximare extrem de puternică și utilă, care stă la baza majorității tehnicilor de analiză și sinteză a vorbirii, inclusiv a celor discutate în acest capitol.

2.2. Raționamentul Extracției de Trăsături

Un sistem de recunoaștere a vorbirii nu poate funcționa eficient dacă procesează direct forma de undă brută a semnalului. Aceasta conține o cantitate imensă de informație, mare parte din ea fiind redundantă sau irelevantă pentru sarcina de recunoaștere. Etapa de extracție a trăsăturilor (feature extraction) este, prin urmare, un pas esențial în orice pipeline ASR, având ca scop transformarea semnalului brut într-o reprezentare mai compactă, mai stabilă și mai informativă din punct de vedere lingvistic.

Trăsăturile acustice ideale trebuie să îndeplinească simultan mai multe criterii:

- **Compactitate:** Să reducă semnificativ dimensionalitatea datelor (de la mii de eșantioane pe secundă la câteva zeci de coeficienți per cadru) fără a pierde informația relevantă.
- **Discriminare:** Să maximizeze separabilitatea între diferite clase fonetice.
- **Robustețe:** Să fie invariante sau cel puțin robuste la variațiile care nu afectează conținutul lingvistic, cum ar fi zgomotul de fond, caracteristicile microfonului (efecte de canal) și particularitățile vorbitorului (ex: pitch).

Cele mai cunoscute și utilizate metode de reprezentare a semnalului vocal sunt coeficienții cepstrali pe scara Mel (MFCC) și codarea liniar predictivă (LPC).

2.3. Coeficienții Cepstrali pe Scara Mel (MFCC)

Coeficienții cepstrali pe scara Mel (MFCC) au devenit standardul de facto în recunoașterea automată a vorbirii de la introducerea lor de către Davis și Mermelstein (1980). Succesul lor se datorează faptului că sunt concepuți pentru a imita particularitățile sistemului auditiv uman, oferind o reprezentare care este atât compactă, cât și relevantă din punct de vedere perceptiv.

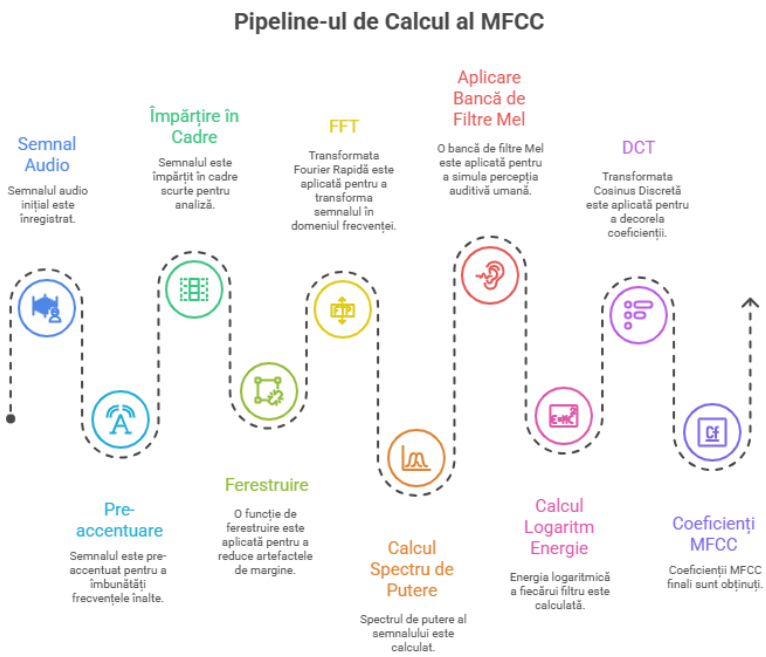

Fig. 1. Etapele calculului coeficientilor MFCC

Calculul MFCC-urilor este un proces multi-etapă (Fayek, 2016), după cum se poate observa și in figura 1:

1. Pre-accentuare (Pre-emphasis): Semnalul vocal are, în general, mai multă energie la frecvențe joase. Pre-accentuarea este un filtru trece-sus care amplifică frecvențele înalte pentru a echilibra spectrul și a îmbunătăți robustețea la zgomot. Se aplică un filtru de ordinul întâi:

$$y[t] = x[t] - \alpha \cdot x[t-1]$$

unde $x[t]$ este semnalul original, iar α este un coeficient de obicei între 0.95 și 0.97.

2. Împărțire în Cadre (Framing): Deoarece vorbirea este un semnal non-staționar, analiza se face pe segmente scurte (cadre), de 20-40 ms, pe durata cărora semnalul este considerat aproximativ staționar. Aceste cadre se suprapun, de obicei cu 10-15 ms, pentru a asigura o tranziție lină între ele.

3. Ferestruire (Windowing): Pentru a reduce discontinuitățile la marginile fiecărui cadru, care ar introduce artefacte în spectrul de frecvență (spectral leakage), fiecare cadru este multiplicat cu o funcție de fereastră, cum ar fi fereastra Hamming:

$$w[n] = 0.54 - 0.46\,cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1$$

unde N este lungimea ferestrei.

4. Transformata Fourier Rapidă (FFT) și Spectrul de Putere: Pentru fiecare cadru ferestruit,

se calculează Transformata Fourier Rapidă (FFT) pentru a obține reprezentarea în domeniul frecvență. Ulterior, se calculează spectrul de putere, care este pătratul magnitudinii rezultatului FFT.

5. Aplicarea Băncii de Filtre Mel: Acesta este pasul cheie care introduce componenta perceptuală. Sistemul auditiv uman nu percepe frecvențele pe o scară liniară; are o rezoluție mai fină la frecvențe joase și mai grosieră la frecvențe înalte. Scara Mel modelează această caracteristică. Conversia de la frecvența în Herți (f) la Mel (m) este dată de:

$$m = 2595 \cdot log_{10}\left(1 + \frac{f}{200}\right)$$

O bancă de filtre triunghiulare (de obicei 20-40 de filtre) este aplicată pe spectrul de putere. Aceste filtre sunt spațiate liniar pe scara Mel, ceea ce înseamnă că sunt mai înguste și mai apropiate la frecvențe joase și mai late și mai distanțate la frecvențe înalte.

6. Calculul Logaritmului Energiei: După ce se calculează energia din fiecare bandă de filtru, se aplică logaritmul. Acest pas este, de asemenea, motivat de percepția umană: nu percepem intensitatea sunetului liniar, ci logaritmic.

7. Transformata Cosinus Discretă (DCT): Energiile logaritmice din băncile de filtre sunt corelate. Pentru a le decorela (o proprietate dorită de multe modele statistice, cum ar fi GMM-urile), se aplică Transformata Cosinus Discretă (DCT). DCT are și proprietatea de a compacta energia în primii coeficienți.

$$c_i = \sum_{j=1}^{M} S_j \backslash cos[\backslash fracπiM(j - 0.5)]$$

unde M este numărul de filtre Mel, S_j este energia logaritmică a filtrului j, iar c_i este al i-lea coeficient cepstral. De obicei, se păstrează doar primii 12-13 coeficienți, deoarece aceștia conțin majoritatea informației despre anvelopa spectrală.

2.4. Codarea Liniar Predictivă (LPC)

O abordare alternativă, cu o istorie bogată în compresia și analiza vorbirii, este Codarea Liniar Predictivă (LPC) (O'Shaughnessy, (2002). Spre deosebire de MFCC, care este motivată de percepție, LPC se bazează direct pe modelul sursă-filtru, încercând să estimeze coeficienții filtrului (tractul vocal).

Ipoteza fundamentală a LPC este că un eșantion al semnalului vocal, $s[n]$, poate fi prezis ca o combinație liniară a celor p eșantioane anterioare (Fellbaum & Richter, 1999):

$$s[n] = -\sum_{k=1}^{p} a_k s[n-k]$$

unde a_k sunt coeficienții de predicție, iar p este ordinul predictorului. Eroarea de predicție, sau semnalul rezidual, este diferența dintre eșantionul real și cel prezis:

$$e[n] = s[n] - \hat{s}[n] = s[n] + \sum_{k=1}^{p} a_k s[n-k]$$

Scopul analizei LPC este de a găsi setul de coeficienți a_k care minimizează eroarea de predicție pe un cadru scurt de semnal. Acești coeficienți a_k formează o reprezentare a filtrului all-pole care modelează cel mai bine anvelopa spectrală a semnalului pentru acel cadru.

LPC a jucat un rol fundamental în dezvoltarea tehnologiilor de compresie a vorbirii (de exemplu, în standardul GSM) și în primele sisteme de sinteză vocală, fiind implementată în dispozitive iconice precum *Speak & Spell*. Deși în ASR modern a fost în mare parte înlocuită de MFCC, LPC rămâne o unealtă valoroasă în analiza fonetică pentru estimarea formanților.

2.5. Tehnici de Normalizare a Trăsăturilor: CMVN

Variațiile introduse de canalul de transmisie (microfon, linie telefonică) sau de particularitățile unui vorbitor pot afecta semnificativ performanța unui sistem ASR. Normalizarea trăsăturilor este o etapă de post-procesare esențială pentru a crește robustețea modelului.

Normalizarea Mediei și Variației Cepstrale (Cepstral Mean and Variance Normalization - CMVN) este cea mai comună tehnică. Obiectivul său este de a transforma vectorii de trăsături astfel încât să aibă o medie de zero și o varianță de unu pe o anumită durată. Acest proces, cunoscut și ca standardizare, elimină deplasările în medie și varianță care pot fi cauzate de caracteristicile constante ale canalului sau ale vorbitorului (Jurafsky and Martin, 2025)

CMVN poate fi aplicat la diferite niveluri:

- **La nivel de enunț (Utterance-level):** Media și varianța sunt calculate și aplicate pentru fiecare enunț în parte. Este simplu de implementat și eficient pentru a elimina variațiile specifice unei singure înregistrări.
- **La nivel de vorbitor (Speaker-level):** Statisticile sunt calculate pe toate enunțurile unui singur vorbitor, fiind util pentru adaptarea la caracteristicile vocale unice ale acestuia.
- **La nivel global (Global):** Media și varianța sunt calculate pe întregul set de date de antrenament și aplicate ulterior la testare.

2.6. Studii de caz și aplicații practice

Numeroase lucrări au demonstrat eficiența acestor tehnici. Milner și Shao (2002) au arătat că semnalul vocal poate fi reconstruit din coeficienții MFCC prin inversarea modelului sursă–filtru, obținându-se o vorbire inteligibilă și apropiată de cea reală. Hsieh și colaboratorii (2016) au propus aplicarea filtrării liniare predictive direct pe seriile temporale cepstrale, obținând o recunoaștere mai robustă în medii zgomotoase. Loweimi și colegii săi (2021) au arătat că separarea explicită a sursei și a filtrului și modelarea acestora prin rețele neuronale convoluționale pot reduce rata de eroare cu până la 14% în sarcini standardizate. Aceste exemple demonstrează valoarea practică și relevanța continuă a metodelor prezentate.

2.7. Concluzii

Modelul sursă–filtru oferă cadrul teoretic necesar pentru înțelegerea și procesarea semnalului vocal. Pe baza acestui model, metodele de extracție a trăsăturilor precum MFCC și LPC transformă semnalul brut într-o reprezentare adecvată pentru algoritmii de recunoaștere și sinteză. MFCC, datorită apropierii sale de percepția auditivă umană, reprezintă standardul actual în recunoașterea automată a vorbirii, în timp ce LPC păstrează o importanță didactică și istorică. În plus, tehnici precum normalizarea cepstrală sporesc robustețea și performanța acestor reprezentări în condiții reale. În ansamblu, fundamentele prezentate în acest capitol constituie bazele pe care se sprijină modelele acustice moderne și aplicațiile inteligente de procesare a vorbirii.