# Recunoasterea vorbitorului

*Groza Gabriel*

## Abstract

Recunoasterea vorbitorului se realizeaza prin extragerea caracteristicilor din secventa vicala urmate de instruirea setului de date si testarea.Pentru extragerea caracteristicilor s-au folosit MFCC si LPC, pentru antrenare s-a folosit algoritmul de cuantizare vectoriala si programul a fost implementat in python.

## 1. Introduction

Recunoasterea vorbitorului este o parte importanta a interactiunii om computer. Este un subiect important in prelucrarea semnalului vocal si are o varietate de aplicatii in special in domeniul securitatii sistemelor.

Semnalele de audiofrecvenţă au spectrul în intervalul 10-20Hz, 20-25kHz si sunt percepute de urechea umană când sunt sub formă de variaţii ale presiunii aerului. Semnalul audio poate fi: vocal sau muzical. Semnalul vocal are spectrul extins de la 20-40 Hz la 8 –10 kHz.[1]

Semnalele vocale au un rol crucial in comunicarea interumana.Sistemul vocal uman este format din totalitatea organelor fona- toare: plamâni, esofag, laringe, corzi vocale,faringe, cavitatea orala, cavitatea nazala, limba, val palatin, maxilar si buze.Avand la baza acest sistem producere a vorbirii,modelarea acestui proces complex necesita anumite simplificari si aproximari, multitudinea proceselor implicate si complexitatea lor neputând fi modelata corect în totalitate.Aplicatiile de codare codare, recunoastere sau sinteza a semnalului vocal pot fi implementate mai usor prin modelarea producerii vorbirii.

Semnalul vocal poate fi inregistrat folosind un microfon. Cu ajutorul diafragmei, microfonul realizeaza conversia miscarii particulelor de aer generate de fonație, în curent electric (semnal). Semnalul astfel rezultat poate fi stocat fie în format analogic, fie digital. Desi în domeniul analogic, informatia stocata este în mod teoretic fara pierderi de informatie, acest format face mai dificila analiza si postprocesarea semnalului înregistrat. Astfel ca, este preferat domeniul digital. [2]

Semnalul audio este cvasi-stationar.Pe intervale temporale scurte poate parea stationar dar semnalul variaza foarte mult in perioade temporale mai mari.Pentru a vizualiza semnalul in frecventa este necesata o transformata Fourier iar reprezentarea 3D a analizei Fourier de durata scurta se numeste spectograma care este o reprezentare atât în timp cât și în frecvență a semnalului.

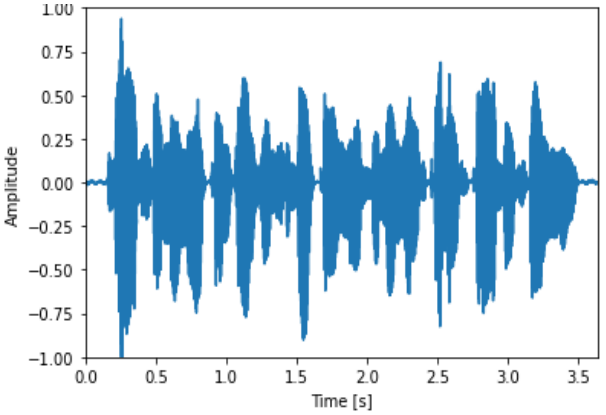


Figura 1.Semnal vocal

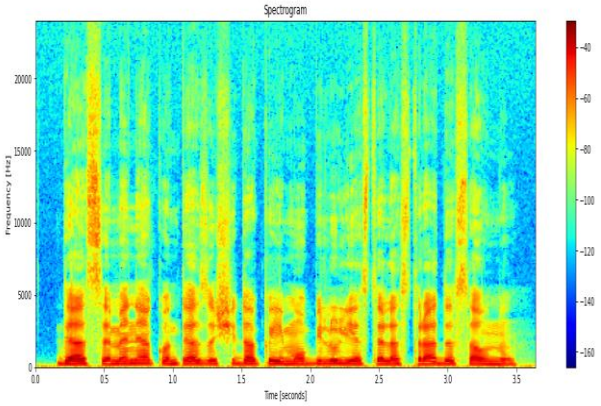


Figura 2.Spectrograma semnalului

## 2. Fundamentare teoretica

Inteligența artificială reprezinta un set de algoritmi care ajuta calculatorul sa generalizeze, sa recunoasca tipare sau caracteristici din date si sa imite intelectul uman. Aceasta, folosind retele neuronale poate lua decizii, care ar necesita in mod normal expertiza umana, utilizand date in timp real.[3]

Retelele neuronale sunt seturi de algoritmi, modelati dupa creierul uman, capabili sa recunoasca tiparele.Pentru ca datele sa fie recunoscute de catre algoritmi acestea trebuie sa fie transformate in date numerice, continute in vectori.[4]

Retelele neuronale sunt utile pentru gruparea si clasificarea datelor in functie de similitudini.Acestea pot extrage si caracteristici provenite de la alti algoritmi pentru grupare si clasificare.

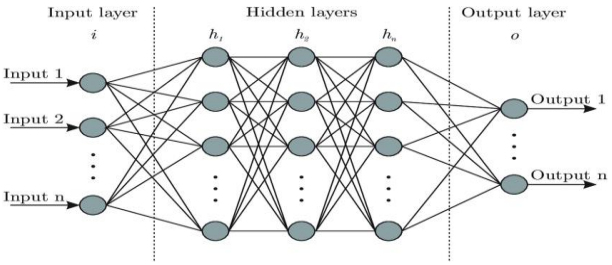


Figura 3. Retea neuronala profunda[5]

Retelele neurolane compuse din mai multe straturi sunt retele profunde.

Straturile sunt formate din noduri, numiti si neuroni, care combina datele de intrare cu un set de coeficienti avand ca rezultat amplificarea sau diminuarea semnificatie intrarii cu privire la sarcina pe care algoritmul incearca sa o invete.Combinatiile de date de intrare si coeficienti sunt insumate si apoi suma este testata pentru a determina daca semnalul respective ar trebui sa treaca in continuare prin reteaua neuronala pentru a contribui la rezultatul final.Daca semnalele trec atunci neuronal a fost acctivat.[6]

Deoarece sunt mult mai profunde, adica au un numar mare de straturi prin care datele trebuie sa treaca in procesul de recunoastere al modelului, retelele de invatare profunda se disting de retelele neuronale cu un singur strat.[6]

Alegerea caracteristicilor de extras din vorbire este cea mai semnificativă parte a recunoașterii vorbitorului. Algoritmul ales utilizeaza in preprocesare MFCC și LPC.

Coeficienții Mel Cepstrali (MFCC) au fost introdusi de Davis și Mermelstein în anii 1980 si sunt o caracteristică utilizată pe scară largă în recunoașterea automată a vorbirii. Acestia reprezinta energia medie din benzile de frecventa date un banc de filtre de lungime N, egal distantate pe scala Mel.In present, analiza MFCC este considerate metoda standard pentru extragerea caracteristicilor din vorbire. [7]

Pasi pentru a calcula coeficientii:

1.Impartim semnalul in cadre de 25ms cu o suprapunere de 10ms. 10ms. Fiecare cadru este multiplicat cu o fereastră Hamming.

2. Periodagrama fiecărui cadru de vorbire este calculată făcând mai întâi un FFT de 512 esantioane pe cadre individuale, luând apoi spectrul de putere:

$$P(k) = \frac{1}{N}|S(k)|^2 \quad (1)$$

3. Întregul interval de frecvență este împărțit în ‘n’ bănci de filtru,care este și numărul de coeficienți pe care îi dorim.

4. Calculam energiile băncii de filtru înmulțind fiecare banc de filtru cu spectrul de putere și adunand coeficienții. Odată realizat acest lucru, rămânem cu „n” numere care ne dau un indicație a cantității de energie din fiecare bancă de filtrare.

5 Luăm logaritmul acestor energii „n” și calculăm transformarea discretă a cosinusului pentru a obține MFCC-urile finale.[8]

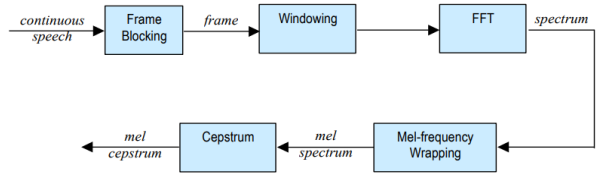


Figura 4.Schema calcularii coeficientilor Cepstral Mel Frequency[8]

Coeficientii de predictie liniara(LPC):

Codarea prin predictive liniara este o metoda digitala pentru codificarea semnalelor analogice in care esantionul curent este prezis ca o combinatie liniare ale ultimelor x esantioane + o erare de predictie.[9]

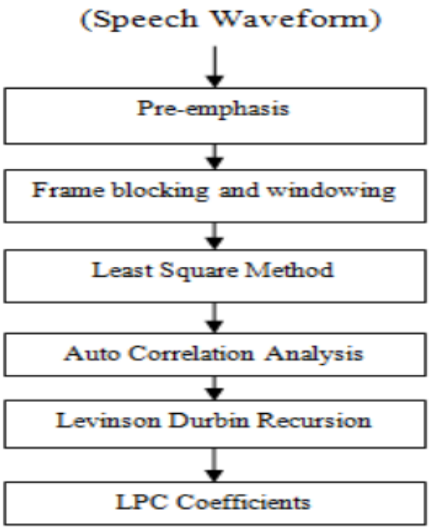


Figura 5. Schema block a procesului de extragere al coeficientilor LPC[9]

Pentru a estima coeficientii LPC, care sunt dati de variabila a, utilizam ecuațiile Yule-Walker care folosec functia de autocorelatie R: [8]

$$R(l) = \sum_{n=1}^{N} x(n)x(n-l) \quad (2)[8]$$

Utilizam forma finala a ecuatiei Yule-Walker :

$$\sum_{k=1}^{p} \alpha_k R(l-k) = -R(l) \quad (3)\ [8]$$

Iar pentru a determina coeficientii prezisi trebuie sa rezolvam un set de p ecuatii cu p necunoscute:

$$\begin{bmatrix} R_0 & R_1 & R_2 \cdots & R_{p-1} \\ R_1 & R_0 & R_1 \cdots & R_{p-2} \\ \vdots & \vdots & \vdots & \vdots \\ R_{p-1} & R_{p-2} & R_{p-3} \cdots & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_p \end{bmatrix} \quad (4)\ [9]$$

In aplicatia practica, pentru a avea rezultate mai precise, coeficientii LPC au fost normalizati astfel incat sa se situeze intre [-1,1].Pentru calcucul coeficientilor “p” ai fiecarui cadru semnalul audio a fost impartit in cadre de 25 ms cu suprapunerea de 10ms.[8]

Cuantizarea vectoriala, folosita initial in compresia datelor este o tehnica de cuantificare clasica folosita in procesarea semnalului care permite modellarea vectorilor de prbabilitate prin distributia vectorilor prototip si este o modalitate foarte eficinta pentru a economisi largimea de banda si spatial de stocare pentru codarea vorbirii. Cuantizarea vectoriala este

folosita pentru compresia datelor cu pierderi si pentru corectarea datelor cu pierderi si estimarea densitatii. Acest algoritm se bazeaza pe invatare competitiva si pe algoritmi de invatare profunda si poate fi inteles ca un cas spcial la unei retele neuronale artificiale.[10]
Cuatizarea vetoriala este un proces de mapare al vectorilor de la un spatiu vectorial mare la un numar finit de regiuni din acel spatiu.Fiecare regiune poarta numele de cluster si este reprezentata de centrul sau numit cuvant de cod. Totalitatea cuvintelor de cod formeaza codebookul.Codbookul este specific si unic pentru fiecare vorbitor si se determina cu ajutorul algortitmului LBG. Algoritmul LGB grupeaza un set de vectori L intr-un set de vecotri de coduri M.Practic algoritmul LGB proiecteaza in mai multe etape un codbook cu M vectori pornid mai intai de la proiectarea unui codbook cu un singur vector apoi foloseste o tehnica de divizare pe cuvinte de cod pentru realiza un cod cu doi vectori si asa mai departe pana ajunge la M-ul dorit [8]

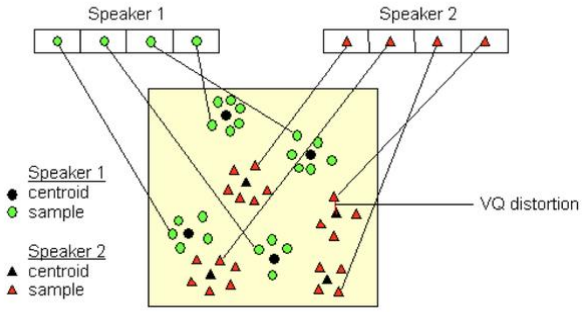


Figura 6.Diagrama conceptuala care ilustraza procesul de cuantizare vectoriala[11]

## 3. Rezultate experimentale

Pentru testare am folosit sase datasetуri diferite
Primul dataset testat este format din 8 inregistrari ale unor femei care pronunta ‘zero’.Acuratetea rezultatelor nu este foarte mare, 37.5% pentru MFCC si 50% pentru LPC.

```
 Training complete
Now speaker  1 features are being tested
Speaker  1  in test matches with speaker  1  in train for training with MFCC
Speaker  1  in test matches with speaker  5  in train for training with LPC
Now speaker  2 features are being tested
Speaker  2  in test matches with speaker  8  in train for training with MFCC
Speaker  2  in test matches with speaker  2  in train for training with LPC
Now speaker  3 features are being tested
Speaker  3  in test matches with speaker  5  in train for training with MFCC
Speaker  3  in test matches with speaker  1  in train for training with LPC
Now speaker  4 features are being tested
Speaker  4  in test matches with speaker  6  in train for training with MFCC
Speaker  4  in test matches with speaker  4  in train for training with LPC
Now speaker  5 features are being tested
Speaker  5  in test matches with speaker  5  in train for training with MFCC
Speaker  5  in test matches with speaker  5  in train for training with LPC
Now speaker  6 features are being tested
Speaker  6  in test matches with speaker  3  in train for training with MFCC
Speaker  6  in test matches with speaker  1  in train for training with LPC
Now speaker  7 features are being tested
Speaker  7  in test matches with speaker  8  in train for training with MFCC
Speaker  7  in test matches with speaker  8  in train for training with LPC
Now speaker  8 features are being tested
Speaker  8  in test matches with speaker  8  in train for training with MFCC
Speaker  8  in test matches with speaker  8  in train for training with LPC
Accuracy of result for training with MFCC is  37.5 %
Accuracy of result for training with LPC is  50.0 %
```

Al 2-lea dataset testat este format din 8 inregistrari audio in care diverse personae pronunta “dogs are sitting by the door”. Acuratetea rezultatelor este de 25% pentru MFCC si 37.5% pentru LPC.

```
Speaker  1  in test matches with speaker  1  in train for training with LPC
Now speaker  2 features are being tested
Speaker  2  in test matches with speaker  1  in train for training with MFCC
Speaker  2  in test matches with speaker  1  in train for training with LPC
Now speaker  3 features are being tested
Speaker  3  in test matches with speaker  1  in train for training with MFCC
Speaker  3  in test matches with speaker  1  in train for training with LPC
Now speaker  4 features are being tested
Speaker  4  in test matches with speaker  6  in train for training with MFCC
Speaker  4  in test matches with speaker  4  in train for training with LPC
Now speaker  5 features are being tested
Speaker  5  in test matches with speaker  1  in train for training with MFCC
Speaker  5  in test matches with speaker  1  in train for training with LPC
Now speaker  6 features are being tested
Speaker  6  in test matches with speaker  6  in train for training with MFCC
Speaker  6  in test matches with speaker  7  in train for training with LPC
Now speaker  7 features are being tested
Speaker  7  in test matches with speaker  6  in train for training with MFCC
Speaker  7  in test matches with speaker  7  in train for training with LPC
Now speaker  8 features are being tested
Speaker  8  in test matches with speaker  1  in train for training with MFCC
Speaker  8  in test matches with speaker  1  in train for training with LPC
Accuracy of result for training with MFCC is  25.0 %
Accuracy of result for training with LPC is  37.5 %
```

Datasetul 3 este format din doua inregistrari in limba araba. Acuratetea rezultatelor este de 50% pentru MFCC si 50% pentru LPC

```
Speaker  1  in test matches with speaker  1  in train for training with MFCC
Speaker  1  in test matches with speaker  1  in train for training with LPC
Now speaker  2 features are being tested
C:\Users\Gabi\Desktop\Speaker-Recognition-master\LPC.py:19: RuntimeWarning: invalid value encountered
in true_divide
  result = r/(variance*(np.arange(n, 0, -1)))
Speaker  2  in test matches with speaker  1  in train for training with MFCC
Speaker  2  in test matches with speaker  1  in train for training with LPC
Accuracy of result for training with MFCC is  50.0 %
Accuracy of result for training with LPC is  50.0 %
```

Al 4-lea dataset testat este format din 8 inregistrari audio mai mari de 10 secunde ale unor femei si barbati.Acuratetea rezultatelor nu este foarte mare, 25% pentru MFCC si 62.5% pentru LPC.

```
Now speaker  1 features are being tested
Speaker  1  in test matches with speaker  6  in train for training with MFCC
Speaker  1  in test matches with speaker  3  in train for training with LPC
Now speaker  2 features are being tested
Speaker  2  in test matches with speaker  6  in train for training with MFCC
Speaker  2  in test matches with speaker  2  in train for training with LPC
Now speaker  3 features are being tested
Speaker  3  in test matches with speaker  6  in train for training with MFCC
Speaker  3  in test matches with speaker  3  in train for training with LPC
Now speaker  4 features are being tested
Speaker  4  in test matches with speaker  1  in train for training with MFCC
Speaker  4  in test matches with speaker  1  in train for training with LPC
Now speaker  5 features are being tested
Speaker  5  in test matches with speaker  2  in train for training with MFCC
Speaker  5  in test matches with speaker  5  in train for training with LPC
Now speaker  6 features are being tested
Speaker  6  in test matches with speaker  6  in train for training with MFCC
Speaker  6  in test matches with speaker  6  in train for training with LPC
Now speaker  7 features are being tested
Speaker  7  in test matches with speaker  2  in train for training with MFCC
Speaker  7  in test matches with speaker  3  in train for training with LPC
Now speaker  8 features are being tested
Speaker  8  in test matches with speaker  8  in train for training with MFCC
Speaker  8  in test matches with speaker  8  in train for training with LPC
Accuracy of result for training with MFCC is  25.0 %
Accuracy of result for training with LPC is  62.5 %
```

Al 5-lea dataset testat este format din 8 inregistrari audio mai mari de 10 secunde ale unor femei.Acuratetea rezultatelor nu este foarte mare, 50% pentru MFCC si 50% pentru LPC.

```
Now speaker  1 features are being tested
Speaker  1  in test matches with speaker  7  in train for training with MFCC
Speaker  1  in test matches with speaker  8  in train for training with LPC
Now speaker  2 features are being tested
Speaker  2  in test matches with speaker  6  in train for training with MFCC
Speaker  2  in test matches with speaker  2  in train for training with LPC
Now speaker  3 features are being tested
Speaker  3  in test matches with speaker  3  in train for training with MFCC
Speaker  3  in test matches with speaker  3  in train for training with LPC
Now speaker  4 features are being tested
Speaker  4  in test matches with speaker  4  in train for training with MFCC
Speaker  4  in test matches with speaker  3  in train for training with LPC
Now speaker  5 features are being tested
Speaker  5  in test matches with speaker  5  in train for training with MFCC
Speaker  5  in test matches with speaker  5  in train for training with LPC
Now speaker  6 features are being tested
Speaker  6  in test matches with speaker  8  in train for training with MFCC
Speaker  6  in test matches with speaker  3  in train for training with LPC
Now speaker  7 features are being tested
Speaker  7  in test matches with speaker  7  in train for training with MFCC
Speaker  7  in test matches with speaker  8  in train for training with LPC
Now speaker  8 features are being tested
Speaker  8  in test matches with speaker  6  in train for training with MFCC
Speaker  8  in test matches with speaker  8  in train for training with LPC
Accuracy of result for training with MFCC is  50.0 %
Accuracy of result for training with LPC is  50.0 %
```

Al 6-lea dataset testat este format din 8 inregistrari audio mai mari de 10 secunde ale unor barbati.Acuratetea rezultatelor nu este foarte mare, 12.5% pentru MFCC si 75% pentru LPC.

```
Now speaker  1 features are being tested
Speaker  1  in test matches with speaker  5  in train for training with MFCC
Speaker  1  in test matches with speaker  1  in train for training with LPC
Now speaker  2 features are being tested
Speaker  2  in test matches with speaker  5  in train for training with MFCC
Speaker  2  in test matches with speaker  2  in train for training with LPC
Now speaker  3 features are being tested
Speaker  3  in test matches with speaker  6  in train for training with MFCC
Speaker  3  in test matches with speaker  2  in train for training with LPC
Now speaker  4 features are being tested
Speaker  4  in test matches with speaker  1  in train for training with MFCC
Speaker  4  in test matches with speaker  4  in train for training with LPC
Now speaker  5 features are being tested
Speaker  5  in test matches with speaker  5  in train for training with MFCC
Speaker  5  in test matches with speaker  5  in train for training with LPC
Now speaker  6 features are being tested
Speaker  6  in test matches with speaker  5  in train for training with MFCC
Speaker  6  in test matches with speaker  6  in train for training with LPC
Now speaker  7 features are being tested
Speaker  7  in test matches with speaker  1  in train for training with MFCC
Speaker  7  in test matches with speaker  4  in train for training with LPC
Now speaker  8 features are being tested
Speaker  8  in test matches with speaker  1  in train for training with MFCC
Speaker  8  in test matches with speaker  8  in train for training with LPC
Accuracy of result for training with MFCC is  12.5 %
Accuracy of result for training with LPC is  75.0 %
```

Pentru o testare suplimentara, folosind datasetul 6 am schimbat numarul de caracteristici MFCC, initial fiind 12 si ordinul coeficientilor LPC, initial fiind 15 in:
nfiltbank =12 si orderLPC =25

```
Accuracy of result for training with MFCC is  12.5 %
Accuracy of result for training with LPC is  62.5 %
```

nfiltbank =12 si orderLPC =17

```
Accuracy of result for training with MFCC is  12.5 %
Accuracy of result for training with LPC is  37.5 %
```

nfiltbank =12 si orderLPC =10

```
Accuracy of result for training with MFCC is  12.5 %
Accuracy of result for training with LPC is  37.5 %
```

nfiltbank =12 si orderLPC =5

```
Accuracy of result for training with MFCC is  12.5 %
Accuracy of result for training with LPC is  50.0 %
```

nfiltbank =7 si orderLPC =4

```
Accuracy of result for training with MFCC is  25.0 %
Accuracy of result for training with LPC is  37.5 %
```

nfiltbank =7 si orderLPC =7

```
Accuracy of result for training with MFCC is  25.0 %
Accuracy of result for training with LPC is  62.5 %
```

nfiltbank =7 si orderLPC =15

```
Accuracy of result for training with MFCC is  25.0 %
Accuracy of result for training with LPC is  75.0 %
```

Observam ca pentru nfiltbank = 7(numarul de caracteristici MFCC) orderLPC =15(ordinul coeficientilor LPC) obtinem o acuratete mai bunba pentru MFCC.

## 4. Concluzii

| Dataset | Acuratete MFCC | Acuratete LPC |
|---|---|---|
| 1 | 37.5% | 50% |
| 2 | 25% | 37.5% |
| 3 | 50% | 50% |
| 4 | 25% | 62.5% |
| 5 | 50% | 50% |
| 6 | 12.5% | 75% |

Se observa per total o acuratete mai mare pentru secventele audio codate cu coeficienti de predictive liniara.

## 5. Bibliografie


[1] http://rf-opto.etc.tuiasi.ro/docs/files/RRCS_cap%202.pdf
[2] Lab#2_WorkingWithSpeechFiles
[3] https://www.scirp.org/html/2-3400378_55265.htm
[4]https://ro.wikipedia.org/wiki/Re%C8%9Bea_neural%C4%3
[5]https://medium.com/@gabriel.mayers/artificial-neural-networks-demystified-d7cbfb6c6916
[6] https://wiki.pathmind.com/neural-network
[7] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/
[8]https://web.stanford.edu/~odas/Documents/speaker_recognition_report.pdf
[9] Lattice Filter Model of Human Vocal Tract Neha Garg1 , Rakesh Garg2 ECE Department, KITM, Kurukeshtra, India
[10] https://en.wikipedia.org/wiki/Vector_quantization
[11] https://timee1994.weebly.com/speaker-recognition.html
[12] https://github.com/orchidas/Speaker-Recognition