Multimodal Machine Learning

Speaker identification

Vasileios Papadopoulos
Dimitrios Delikonstantis

September 17, 2021

Motivation

- Speaker identification is the process of determining from which of the registered speakers a given utterance comes
- Speaker identification (in conjunction with speaker verification) is widely used in security systems
- The task of identifying a person by his voice becomes increasingly crucial with the development of IoT and technology
- Many applications require speaker identification
    - Biometrics authentication
    - Voice mail
    - Smart home

Dataset

VoxCeleb audio dataset

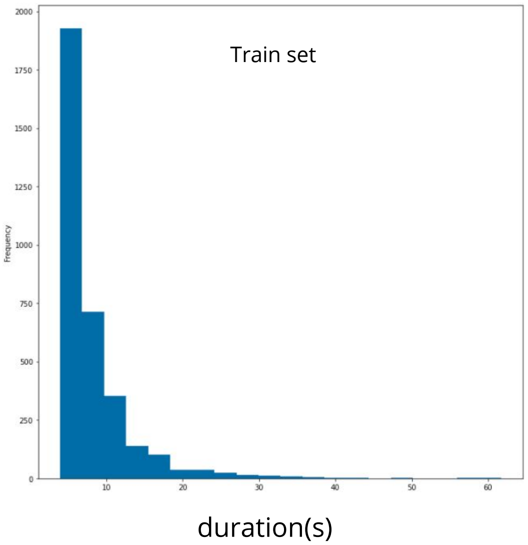
- Contains over 100,000 human speech utterances for 1,251 celebrities
- Spanning a wide range of different ethnicities, accents, professions and ages
- We focused only on **39** speakers with total of **4837** audio clips(computational issues)

Train/Test/Split

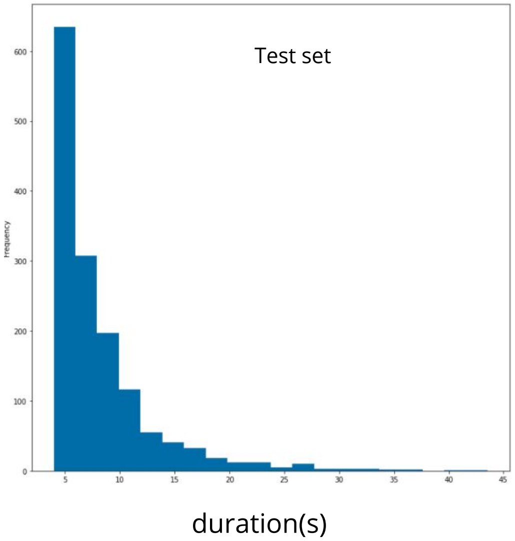
- For each speaker:
  - 70% audio clips kept for training
  - 30% testing
- Training set:
  - 70% train
  - 30% validation

Duration distribution

- Majority of clips have duration up to 10s.
- Sampling rate is 16KHz



- Majority of clips have duration up to 10s.
- Sampling rate is 16KHz

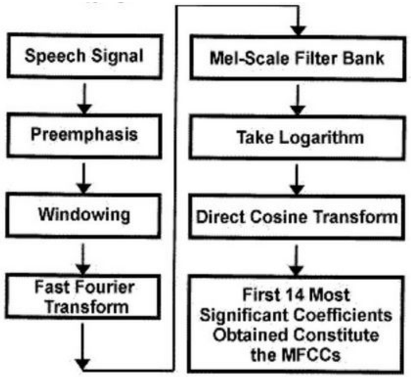


Shortest clip: **3.9(s)**

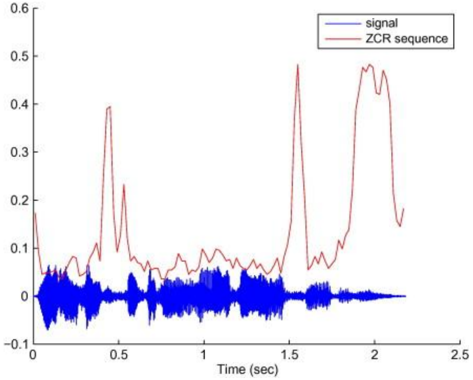
We use the first **2s** of each clip.
Assumption: no silence at the beginning.

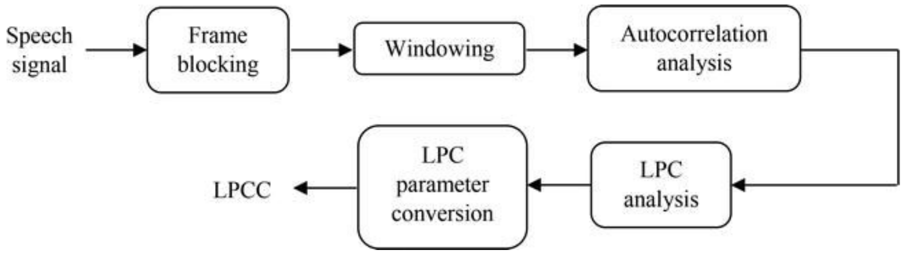
Feature extraction

- Mel-frequency cepstral coefficients (MFCCs)
  - Short-term representation power spectrum of a sound
- Zero-crossing rate (ZCR)
  - Rate at which a signal changes its sign from positive to negative or vice versa
- Linear Prediction Coefficients (LPC)
  - Future values of a discrete-time signal are estimated as a linear function of previous samples



Speech Signal → Preemphasis → Windowing → Fast Fourier Transform → Mel-Scale Filter Bank → Take Logarithm → Direct Cosine Transform → First 14 Most Significant Coefficients Obtained Constitute the MFCCs



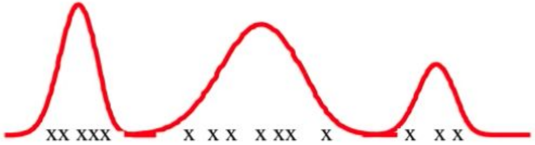


Speech signal → Frame blocking → Windowing → Autocorrelation analysis → LPC analysis → LPC parameter conversion → LPCC

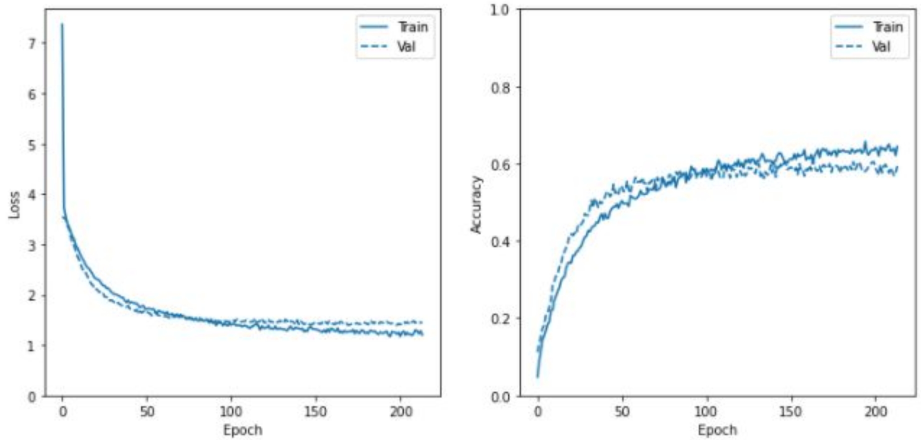
Models

- “Deep” Neural Network (113K parameters)
    - 5 Dense layers
    - 0.3 dropout
    - Softmax
    - Categorical_crossentropy
    - Adam optimizer
    - Early stopping
- Gaussian Mixture Models
    - For each speaker Σ(<mfcc,zcr,lpc>) (cluster) model as Gaussian Distribution
        - Not just by their mean(K-means)
    - Gives probability model of X
    - Perform statistical inference
        - Assign data to cluster with some probability

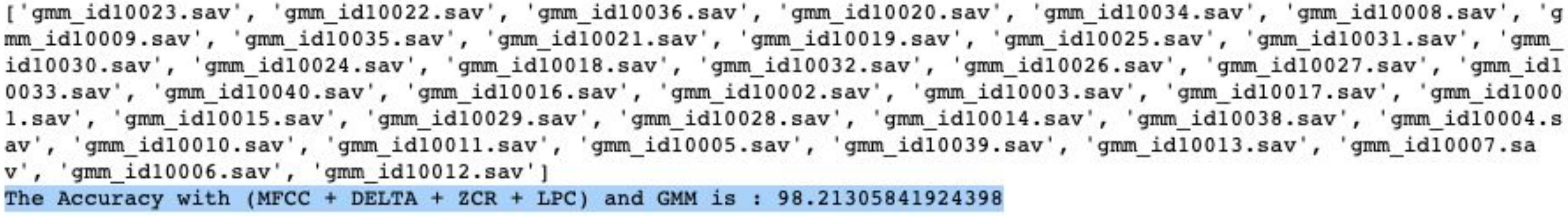
Practically, a multi-class problem with 39 classes(speakers)



Results



- Model underfits.
- Too simple.
- Deeper architecture required



```
['gmm_id10023.sav', 'gmm_id10022.sav', 'gmm_id10036.sav', 'gmm_id10020.sav', 'gmm_id10034.sav', 'gmm_id10008.sav', 'g
mm_id10009.sav', 'gmm_id10035.sav', 'gmm_id10021.sav', 'gmm_id10019.sav', 'gmm_id10025.sav', 'gmm_id10031.sav', 'gmm_
id10030.sav', 'gmm_id10024.sav', 'gmm_id10018.sav', 'gmm_id10032.sav', 'gmm_id10026.sav', 'gmm_id10027.sav', 'gmm_id1
0033.sav', 'gmm_id10040.sav', 'gmm_id10016.sav', 'gmm_id10002.sav', 'gmm_id10003.sav', 'gmm_id10017.sav', 'gmm_id1000
1.sav', 'gmm_id10015.sav', 'gmm_id10029.sav', 'gmm_id10028.sav', 'gmm_id10014.sav', 'gmm_id10038.sav', 'gmm_id10004.s
av', 'gmm_id10010.sav', 'gmm_id10011.sav', 'gmm_id10005.sav', 'gmm_id10039.sav', 'gmm_id10013.sav', 'gmm_id10007.sa
v', 'gmm_id10006.sav', 'gmm_id10012.sav']
The Accuracy with (MFCC + DELTA + ZCR + LPC) and GMM is : 98.21305841924398
```

Demo

September 17, 2021

Conclusions

- We combined 3 widely used features for speaker identification.
    - GMM performed well on tiny dataset
    - NN-tuning required
- The combination of features seem to work for identification tasks but it is unclear(to us) whether could be useful for ‘similarity’ tasks.

Future work

- Work with entire vox celeb dataset
- Measure euclidean distance of features in hyperspace
    - Assumption: similar speakers will be closer
- Measure cosine similarity of features in hyperspace
- Type of Auto-encoder
- Combination of different feature extraction methods