Deep Neural Networks for Speaker Recognition

Xingxing Tang

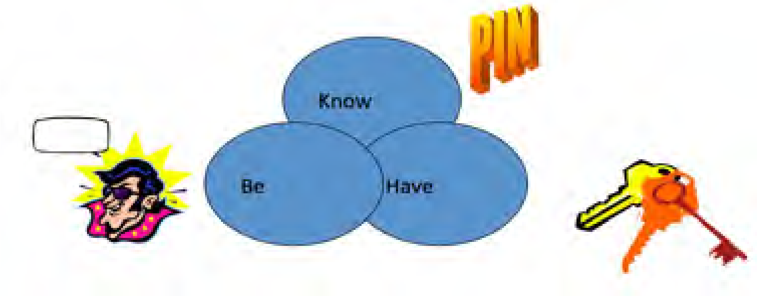
March 28, 2017

Outline

- Introduction
- Algorithm
  - VAD
  - Spectrogram
  - Model structure
  - Triplet Loss
- Results
- References

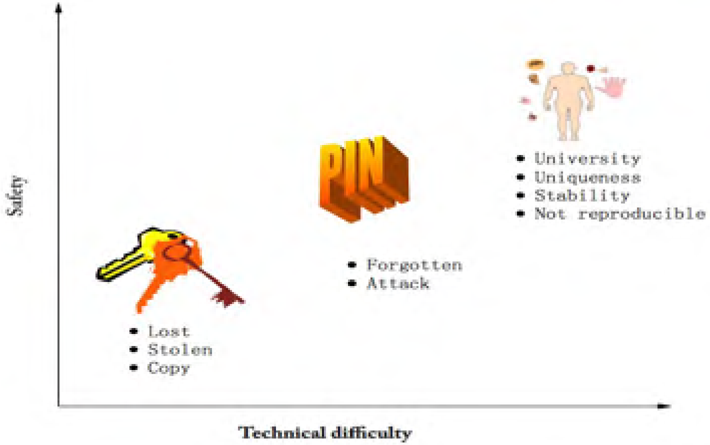
What is Biometrics?

The automated use behavioral and physiological characteristics to determine or veiry an identity.



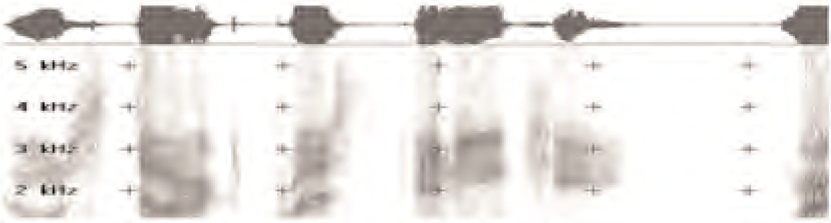
Biometrics

Comparison of various types of authentication technology



Human Speech

Is my voice really unique?There is only one you!Your voice print consists of physical characteristics such as your nasal passages, vocal cords, cadence of speech and duration of your vocal pattern. These characteristics are measured and then turned into a digital format called a spectrogram as illustrated below.



Speaker Recognition

Speaker recognition is the identification of the person who is speaking by characteristics of their voices

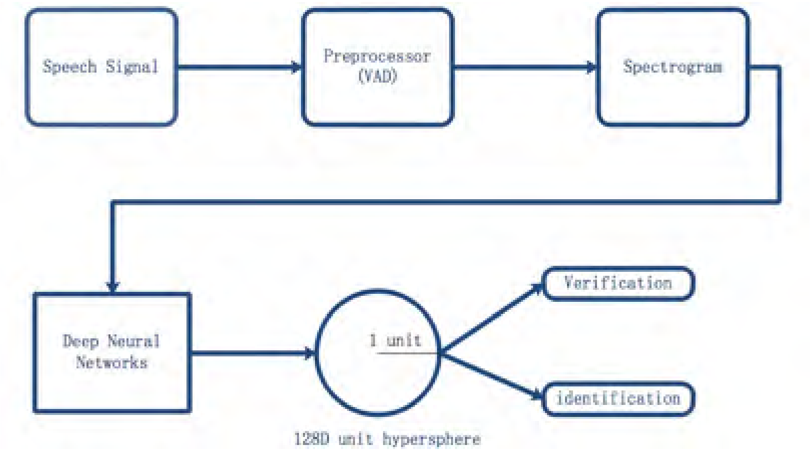
- Speaker identification is the task of determining which speaker
- Speaker verification aims at accepting or rejecting the identity claim

Variants of Speaker Recognition

Speaker recognition systems fall into two categories : text-dependent and text-independent.

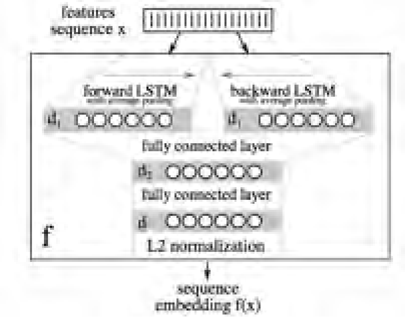
- Text-Dependent
  - Limited speak text
  - Language related
- Text-Independent
  - Not limit speak text
  - Not language relate
  - Language related

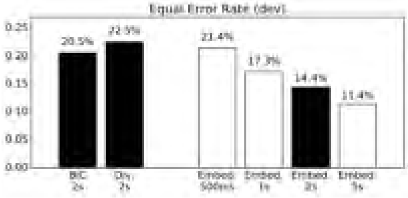
Pipeline



Speech Signal
Preprocessor (VAD)
Spectrogram
Deep Neural Networks
1 unit
Verification
identification
128D unit hypersphere

Tristounet



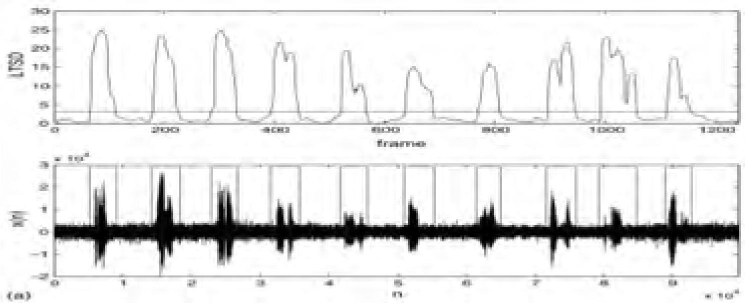


[Herve Bredin]

VAD-LTSD

Voice Activity Detection shall be applied for all signals as a prefilter,We use LTSD(Long-Term Spectral Divergence) algorithm.

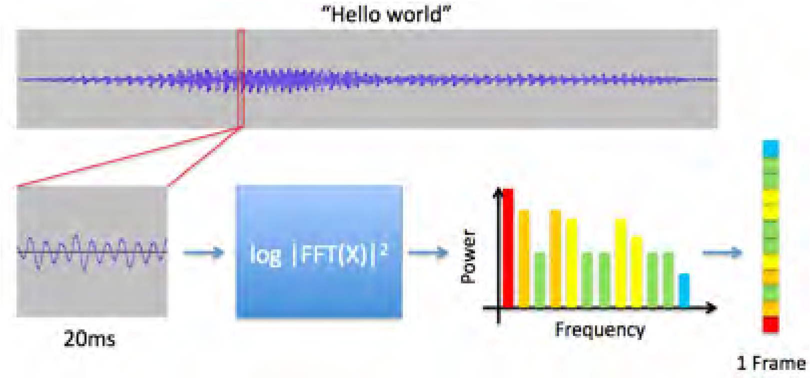
- Splits a utterance into overlapped frames and give scores for each frame on the probability that there is voice activity in this frame.
- Accumulated to extract all the intervals with voice activity.



Spectrogram

Take a small window of waveform.

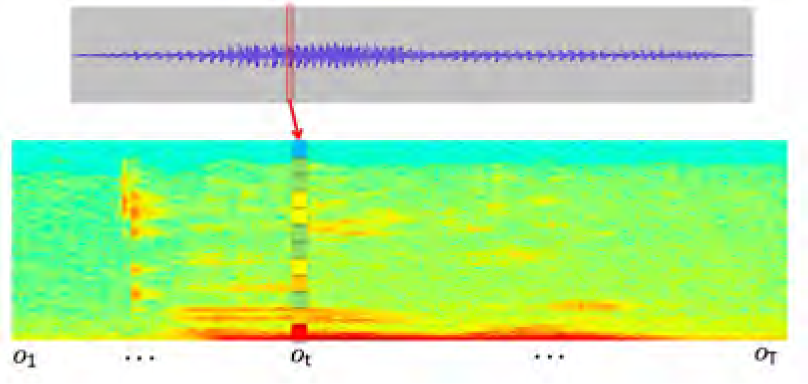
- Compute FFT and take magnitude.(i.e.,power)
- Describes frequency content in local window



[Adam Coates et al.]

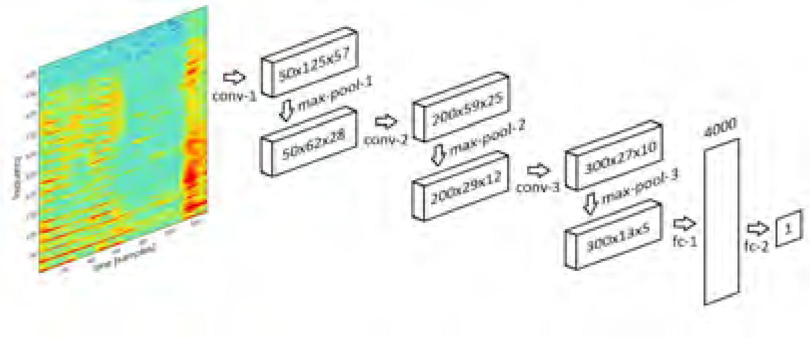
Spectrogram

Concatenate frames from adjacent windows.



[Adam Coates et al.]

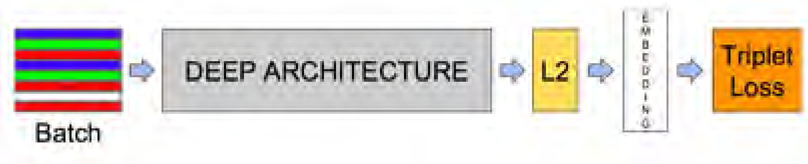
Deep Neural Networks



[Marek Hruzet al.]

Model Structure

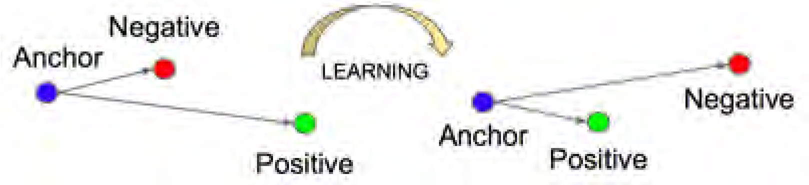
A Deep CNN followed by L2 normalization,During training this is followed by the triplet loss.



[Florian Schroff et al.]

Triplet Loss

The Triplet Loss minimizes the distance between an anchor and a positive, both of which have the same identity, and maximizes the distance between the anchor and a negative of a different identity.



[Florian Schroff et al.]

Triplet Loss

The triplet loss is motivated by

$$\left|\left|f(x_i^a) - f(x_i^p)\right|\right|_2^2 + \alpha < ||f(x_i^a) - f(x_i^n)||_2^2 \tag{1}$$

where α is a margin that is enforced between positive and negative pairs.

The loss that is being minimized is then $L =$

$$\sum_{i=1}^{n}[\left|\left|f(x_i^a) - f(x_i^p)\right|\right|_2^2 + \alpha - ||f(x_i^a) - f(x_i^n)||_2^2] \tag{2}$$

In order to ensure fast convergence it is crucial to select triplets that violate the triplet constraint in Eq. (1).

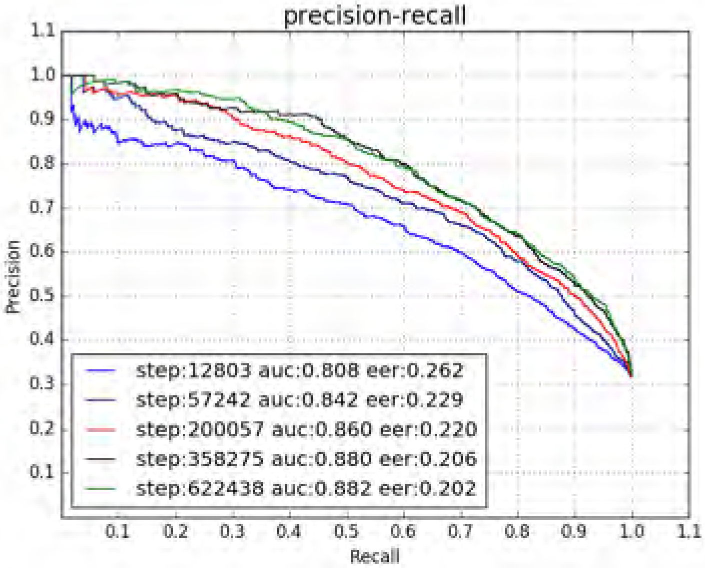
Evaluation Metric

Two types of errors exist: false positive and false negative

- False Positive
  - Two sequences from two different speakers are incorrectly classified as uttered by the same speaker
- False Negative
  - Two sequences from the same speaker are classified as uttered by two different speakers

The higher (resp. lower) the decision threshold is, the higher the false negative (resp. positive) rate is (FNR, FPR). We report the equal error rate (EER), i.e. the value of FPR and FNR when they are equal.

Result



precision-recall
Precision
Recall
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
step:12803 auc:0.808 eer:0.262
step:57242 auc:0.842 eer:0.229
step:200057 auc:0.860 eer:0.220
step:358275 auc:0.880 eer:0.206
step:622438 auc:0.882 eer:0.202

References

- Javier Ramirez,et al.:Efficient voice activity detection algorithms using long-term speech information
- Dario Amodei,et al.:Deep Speech 2: End-to-End Speech Recognition in English and Mandarin
- Florian Schroff, Dmitry Kalenichenko, James Philbin:FaceNet: A Unified Embedding for Face Recognition and Clustering
- Xinyu Zhou,Yuxin Wu,Tiezheng Li:Digital Signal Processing:Speaker Recognition Final Report
- Adam Coates,Vinay Rao:Speech Recognition and Deep Learning
- Marek Hruz,Marie Kunesova:Convolutional Neural Network in the Task of Speaker Change Detection
- Herve Bredin:TRISTOUNET: TRIPLET LOSS FOR SPEAKER TURN EMBEDDING

Company

xingtang

北京 海淀

扫一扫上面的二维码图案，加我微信