Voice Biometrics

Phan Trung Kiên - Phòng XLTN – Khối CNTLA
18/10/2024

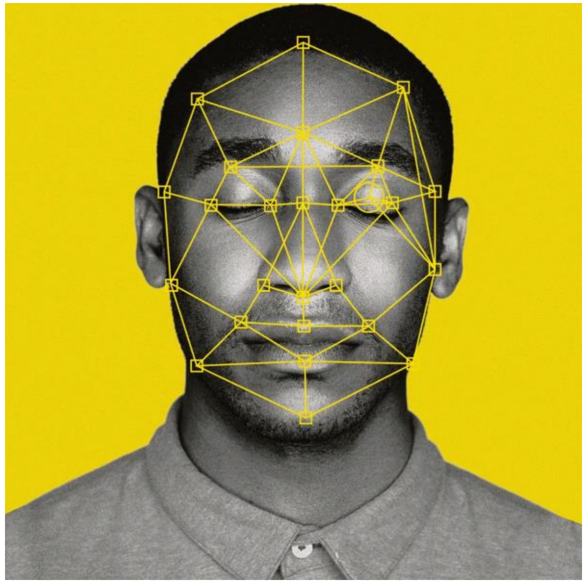
I. **Introduce**

II. Speaker Recognition

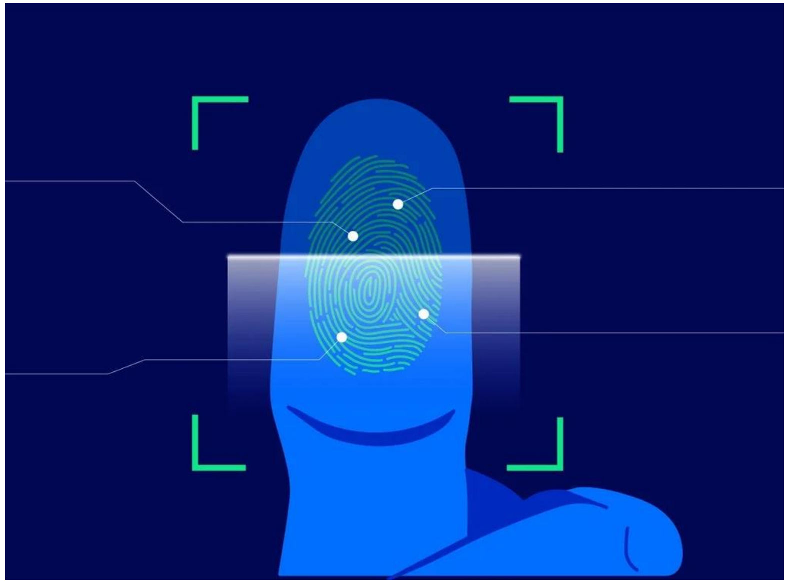
III. Speaker Classification

Biometrics definition

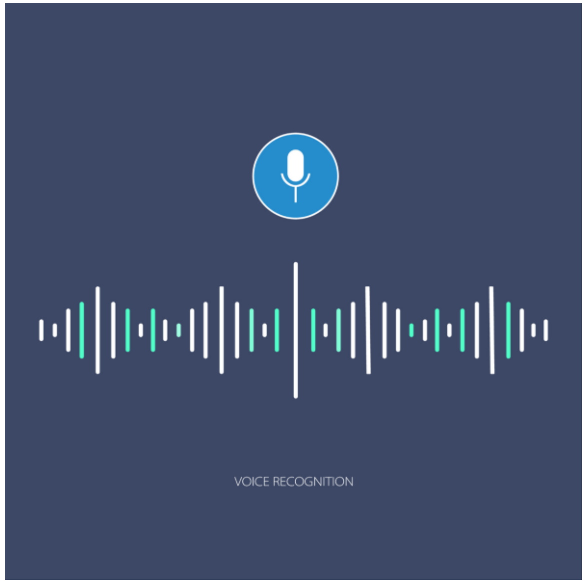
Unique **physical** or **behavioral** human characteristics that can be used to digitally identify a person to grant access to systems, devices or data.



Face ID



Fingerprint



Voiceprint

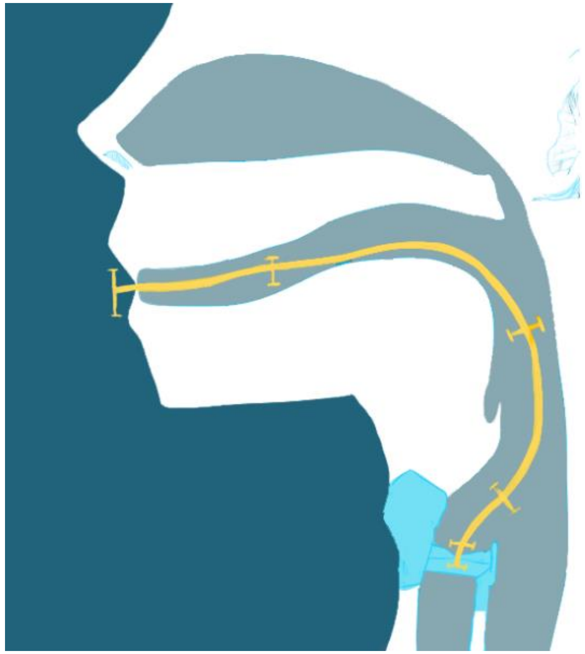
Voiceprint

Behavioral and Physical characteristic factors are combined to produce a unique voice pattern for each individual.

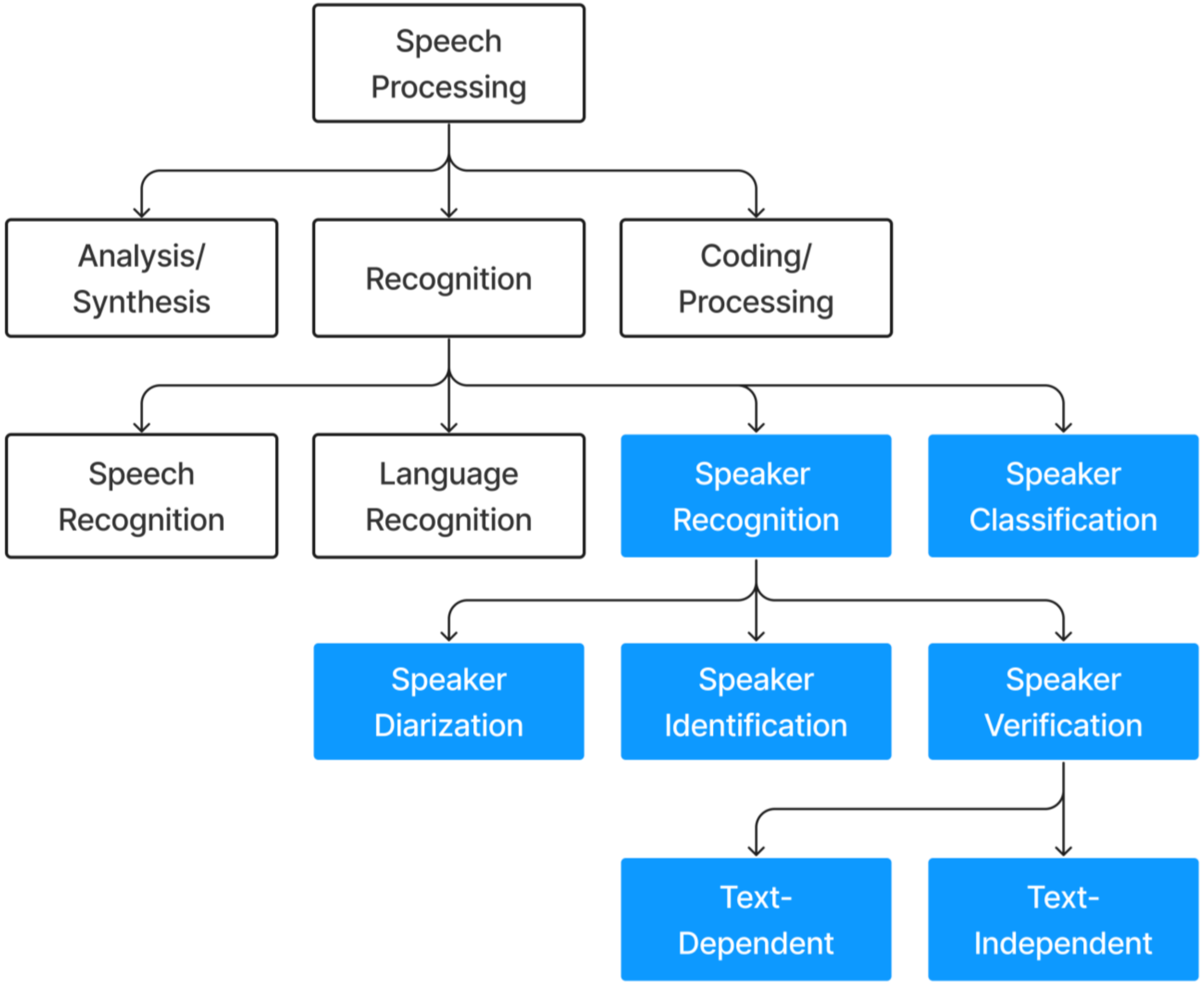
Biometric technology captures this pattern as a voiceprint.

- Speed of Speech
- Pronunciation and Emphasis
- Accents

- Unique Physical Traits of Vocal Tract
- Mouth Shape and Size
- Nasal Passages



Voice Biometrics

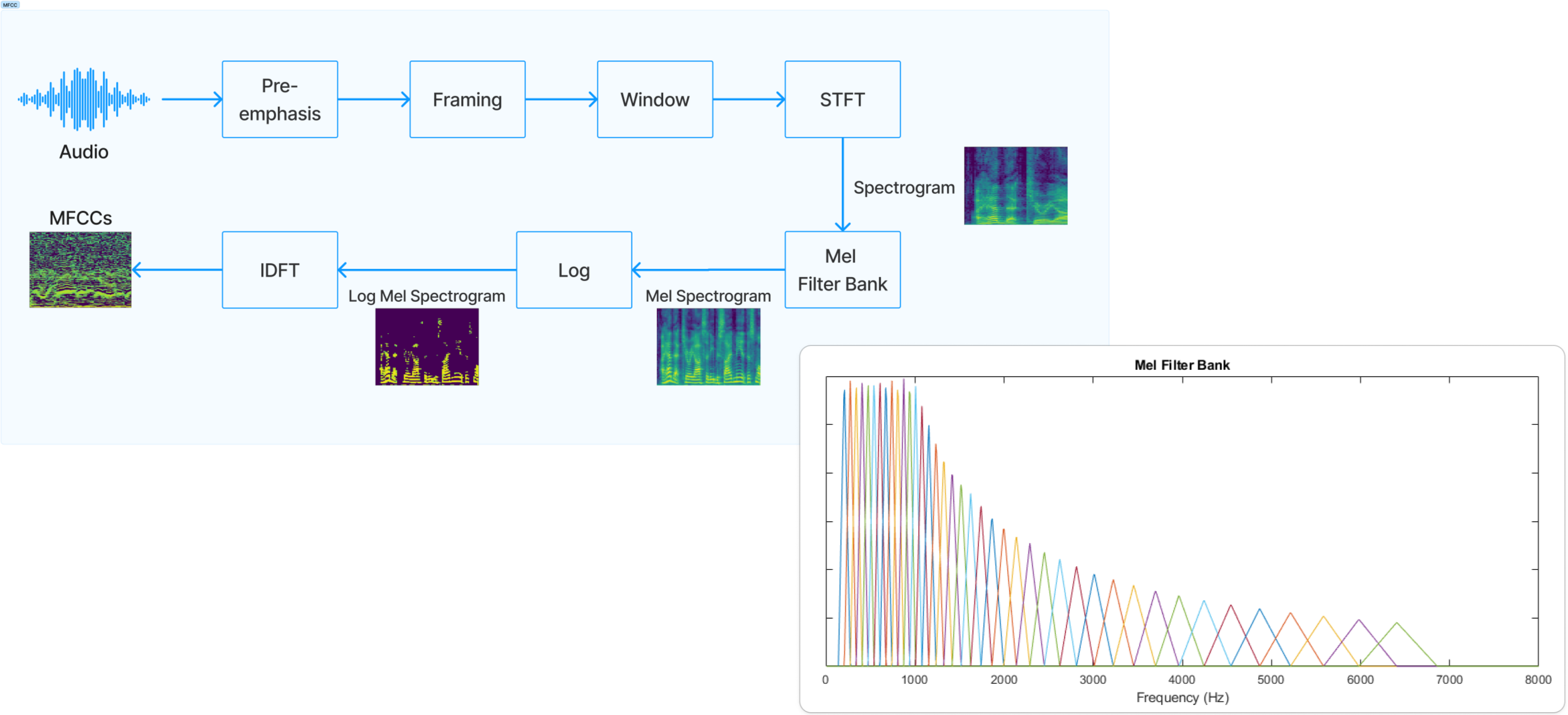


I. Introduce

II. Speaker Recognition

III. Speaker Classification

Speech preprocessing



MFCC

Audio → Pre-emphasis → Framing → Window → STFT

Spectrogram

Mel Filter Bank

Mel Spectrogram

Log

Log Mel Spectrogram

IDFT

MFCCs

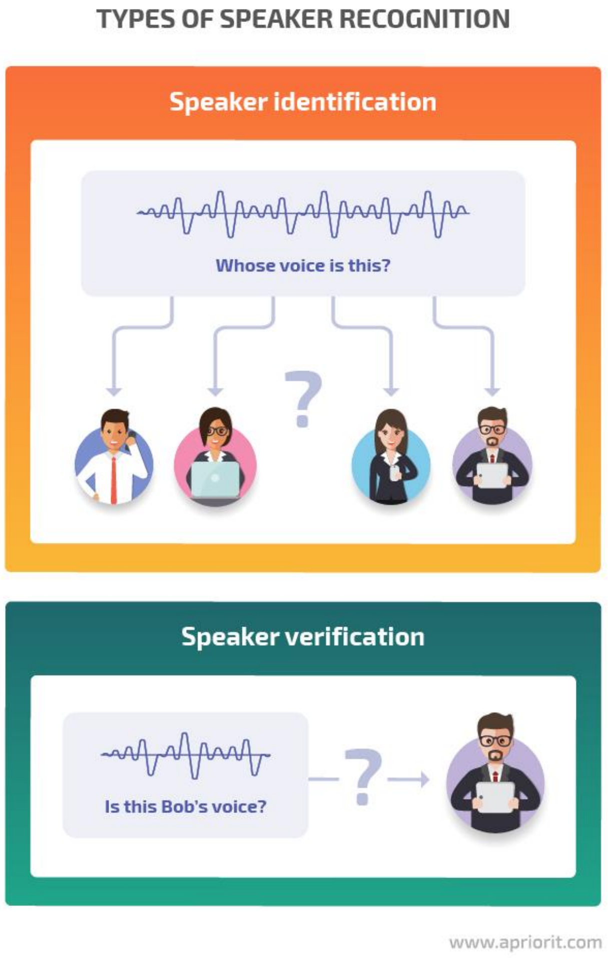
Speaker Identification & Verification

Speaker Verification:

- The speaker claims to be of a certain identity and the voice is used to verify this claim.
- A 1:1 match where one speaker's voice is matched to a particular template.

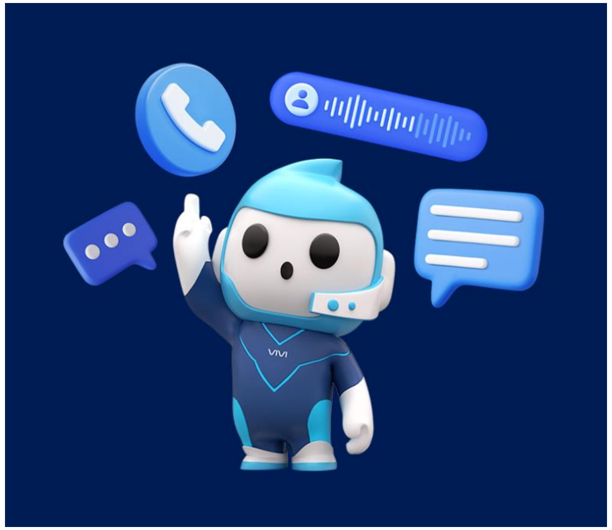
Speaker Identification:

- Determining an unknown speaker's identity.
- A 1:N match where the voice is compared against multiple templates.

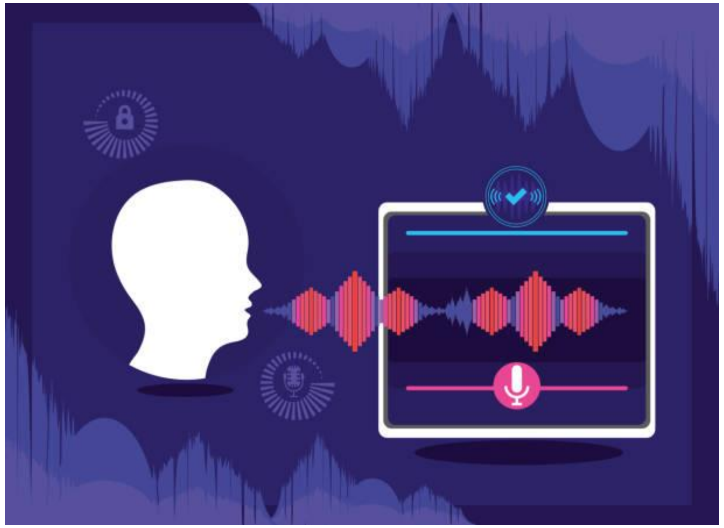


www.apriorit.com

Applications



Call-bot



Authentication

Voice Assistant

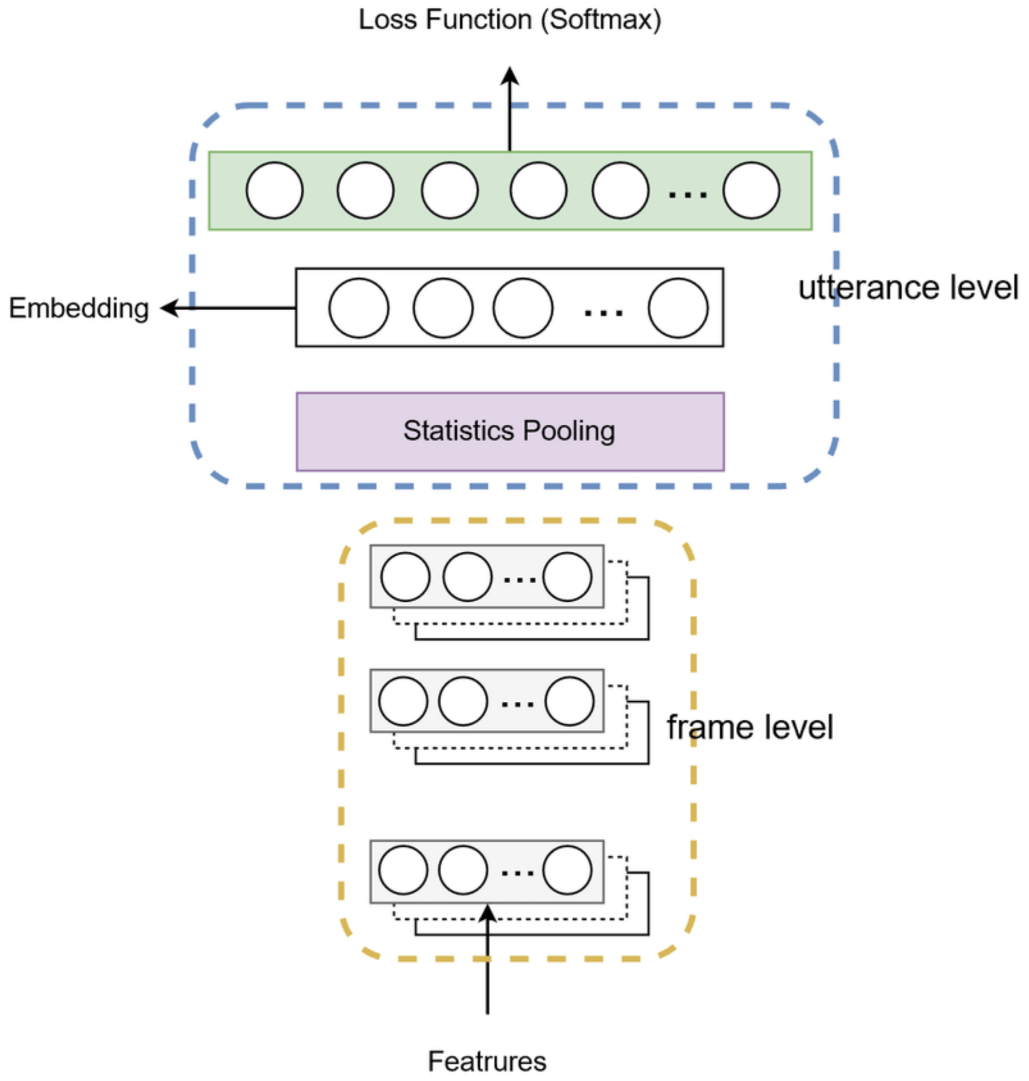
Speaker Embedding

Model backbone:

- ResNet
- TDNN
- CNN-TDNN
- ECAPA-TDNN

Pooling:

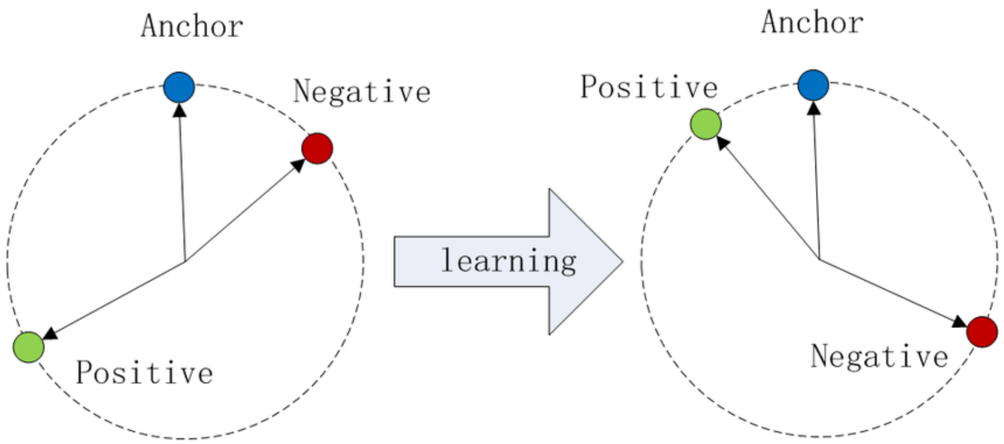
- Statistics Pooling
- Attentive Statistics Pooling
- Multi-Head Attention Pooling

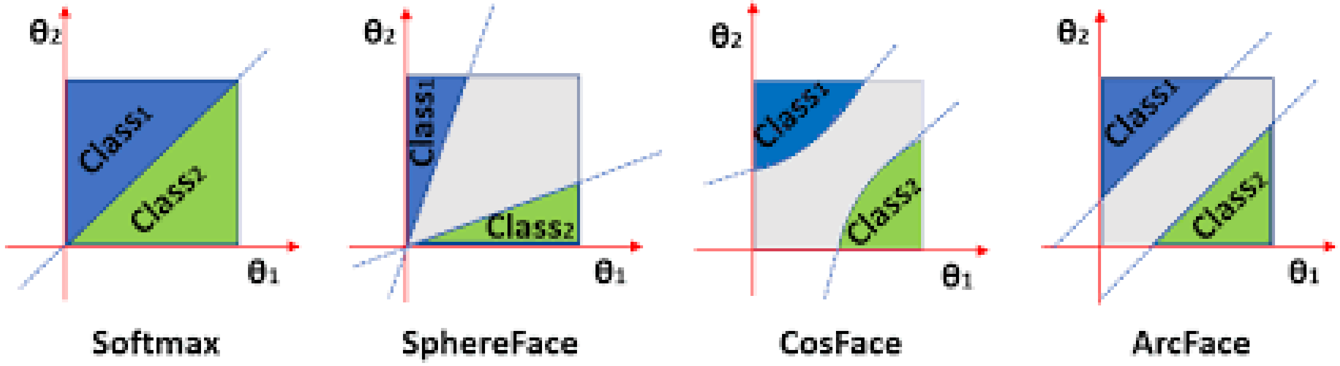


Speaker Embedding

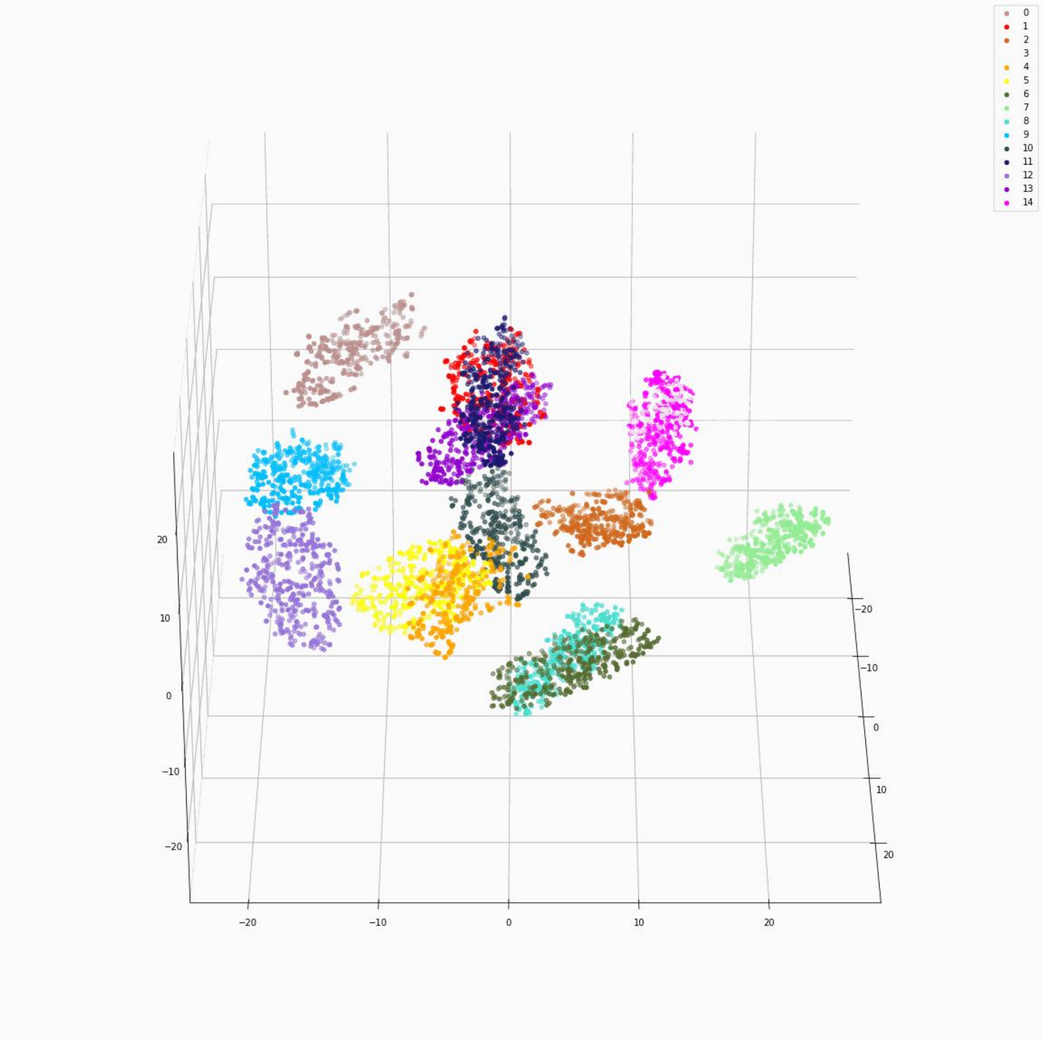
Loss Function:

- Metric Loss:
  - Triplet Loss
  - Contrastive Loss
- Classification Loss:
  - Softmax Loss
  - A-Softmax Loss
  - AM-Softmax Loss
  - AAM-Softmax Loss

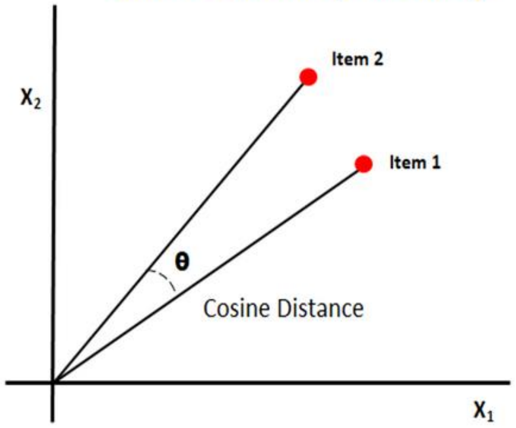




Speaker Embedding



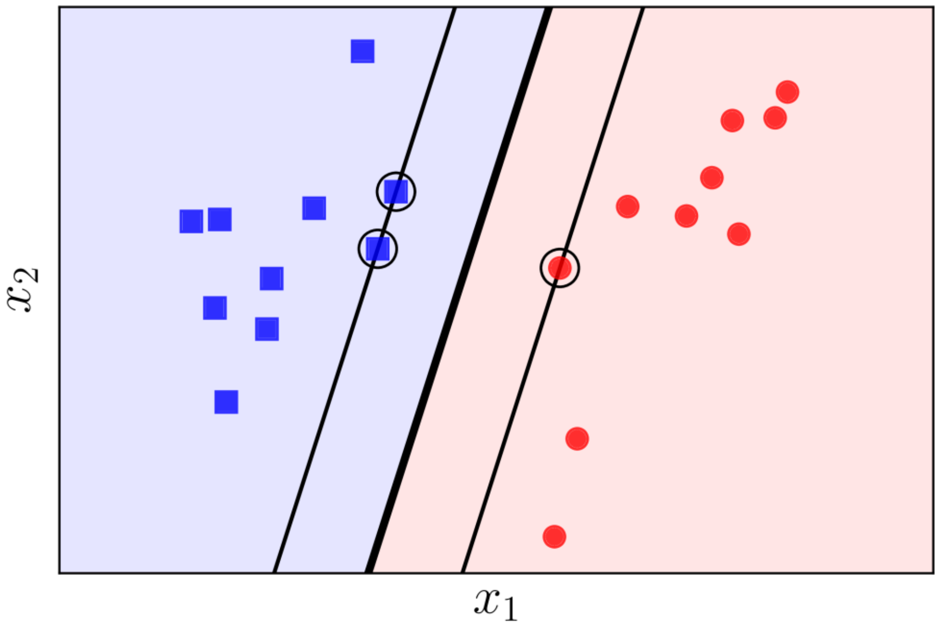
Cosine Distance/Similarity

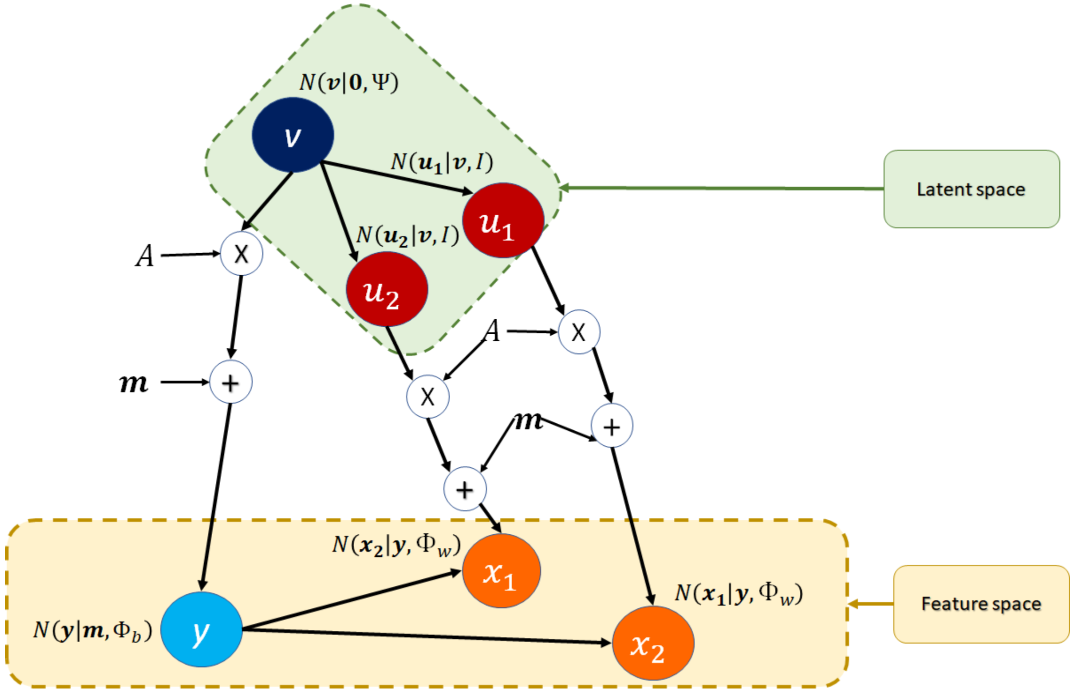


$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

Back-end

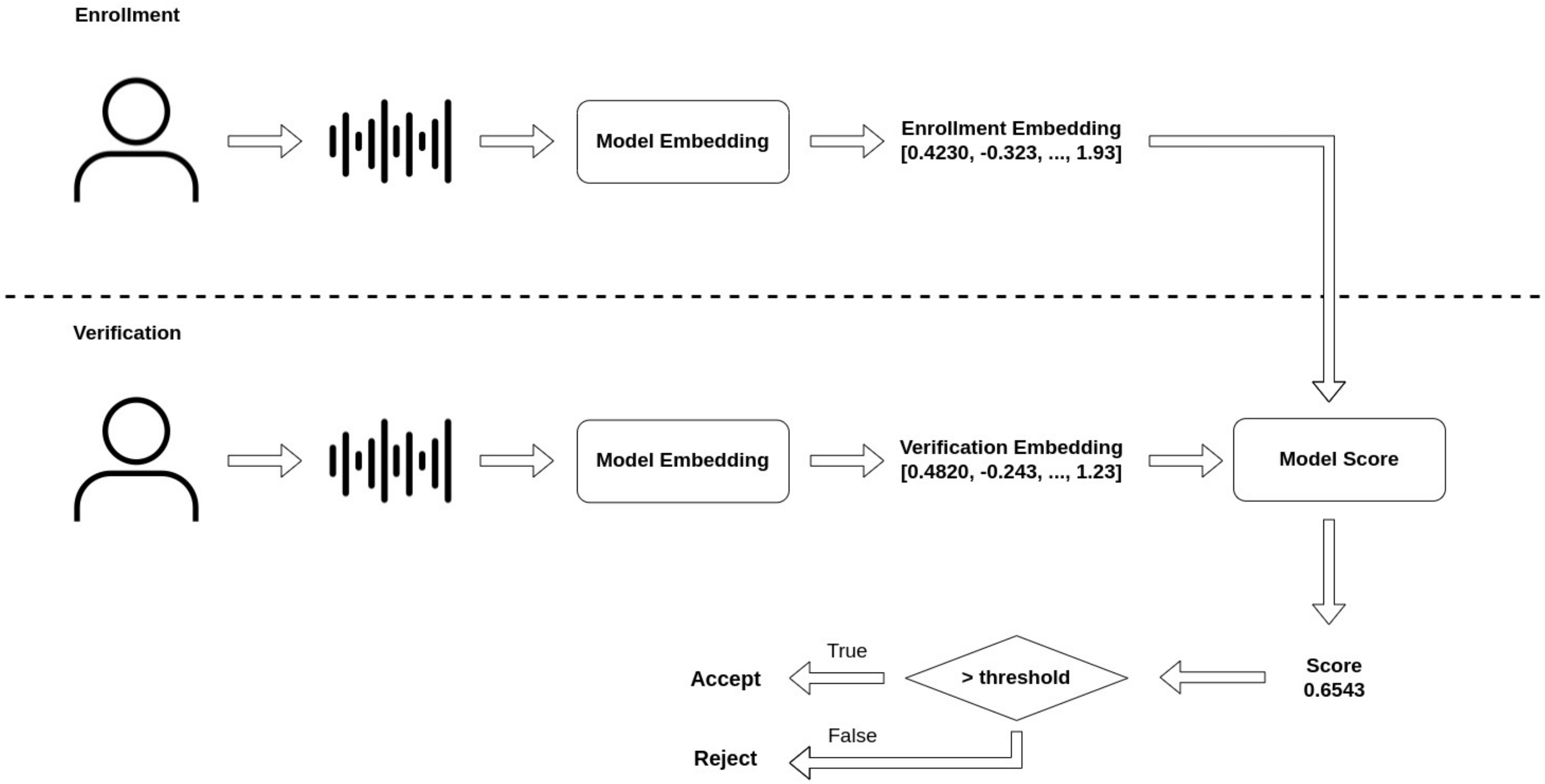
- Cosine Similarity
- Basic Classifiers: SVM, GMM, Logistic Regression (LR)
- PLDA Classifiers: PLDA, APLDA, CORAL, …



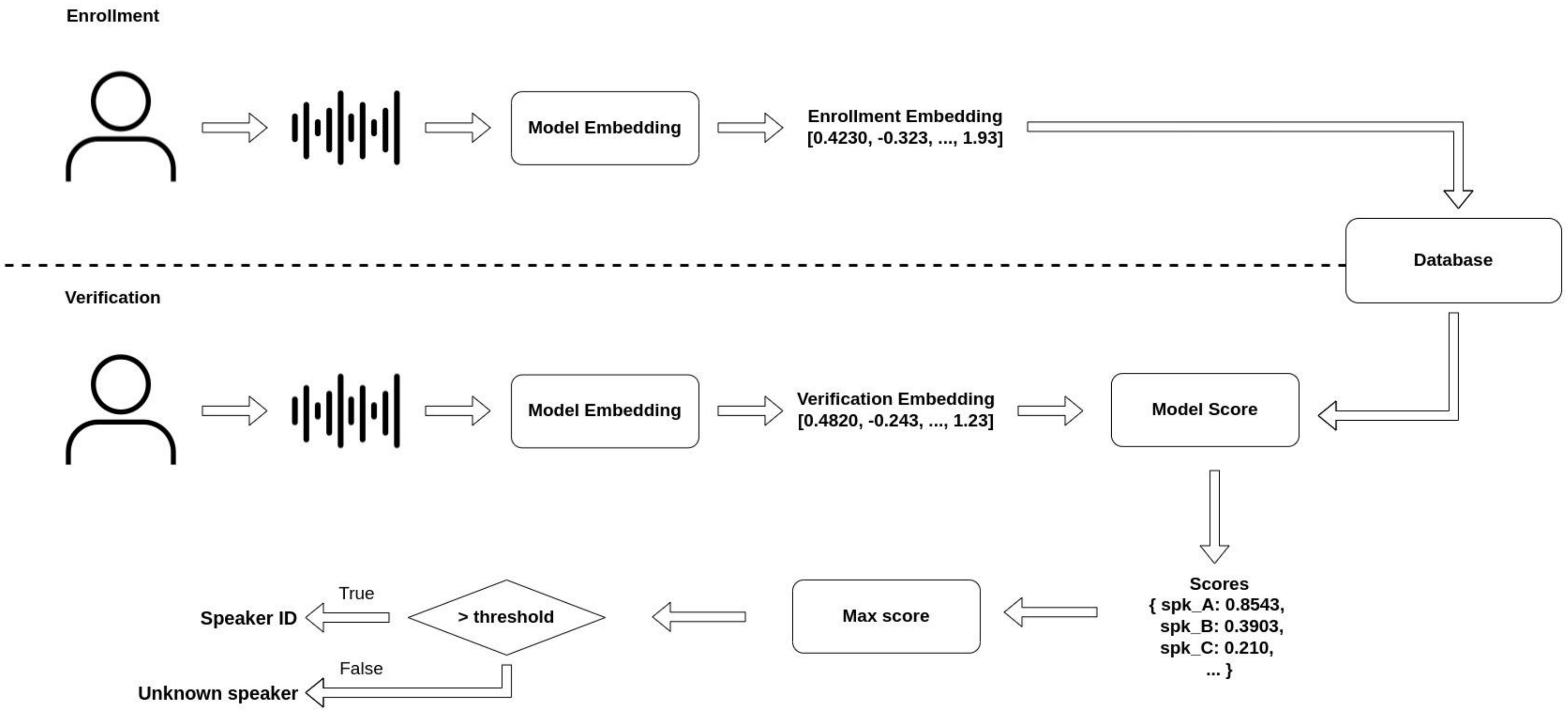


$N(\boldsymbol{v}|\boldsymbol{0},\Psi)$
v
$N(\boldsymbol{u_1}|\boldsymbol{v},I)$
u_1
$N(\boldsymbol{u_2}|\boldsymbol{v},I)$
u_2
A
X
$\boldsymbol{m}$
+
A
X
X
$\boldsymbol{m}$
+
+
Latent space
$N(\boldsymbol{x_2}|\boldsymbol{y},\Phi_w)$
x_1
$N(\boldsymbol{x_1}|\boldsymbol{y},\Phi_w)$
x_2
$N(\boldsymbol{y}|\boldsymbol{m},\Phi_b)$
y
Feature space

Speaker Verification



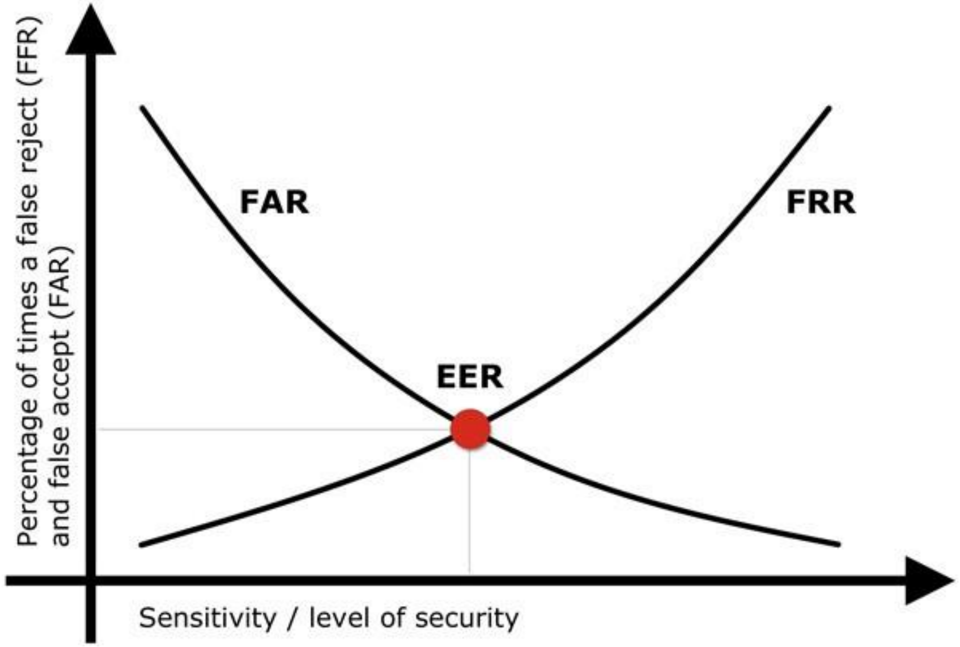
Speaker Identification



Metrics

Equal Error Rate & Minimum Detection Cost Function

EER



Min DCF

$$C_{\mathrm{det}}(P_{\mathrm{miss}}, P_{\mathrm{FA}}) = C_{\mathrm{miss}} P_{\mathrm{miss}} P_{\mathrm{tar}} + C_{\mathrm{FA}} P_{\mathrm{FA}} (1 - P_{\mathrm{tar}})$$

C_{miss} - cost of a miss target (false reject)
C_{FA} – cost of a false alarm (false accept)
P_{miss} and P_{FA} are determined by the evaluator by counting errors.
P_{tar} is the prior probability that a target speaker event occurs in the application.

Challenges

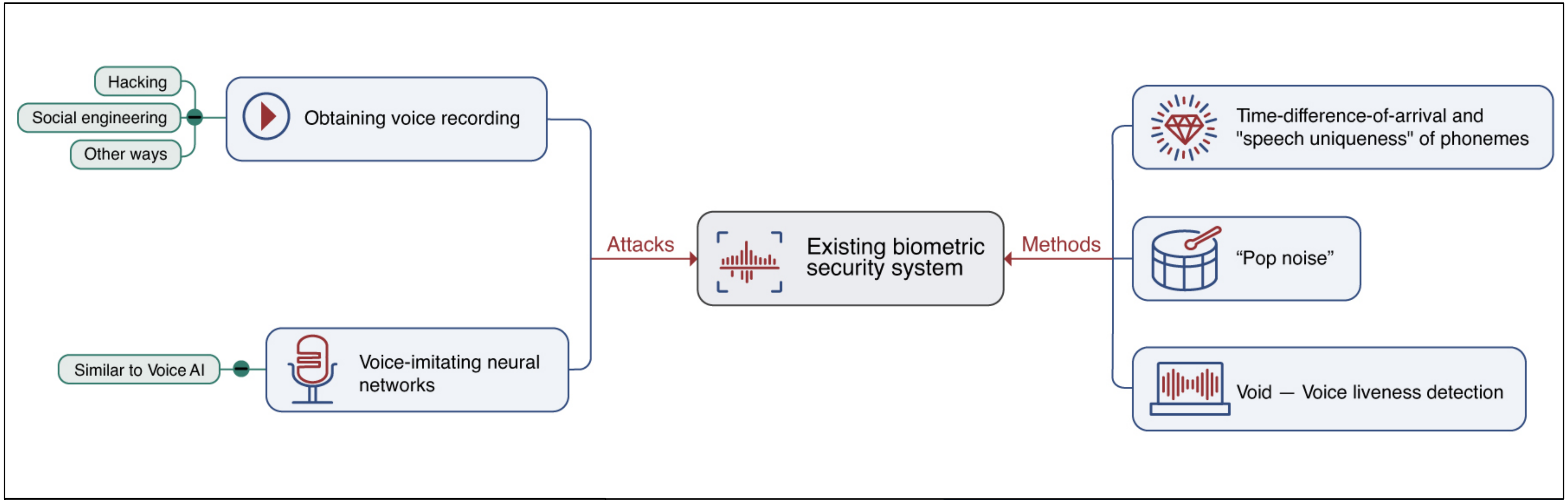
Data

Voice Quality

Spoofing

Cross Device

Voice anti-spoofing


Hacking
Social engineering
Other ways
Obtaining voice recording
Similar to Voice AI
Voice-imitating neural networks
Attacks
Existing biometric security system
Methods
Time-difference-of-arrival and "speech uniqueness" of phonemes
"Pop noise"
Void — Voice liveness detection

Logical access:

- TTS
- Voice Conversion

Physical access:

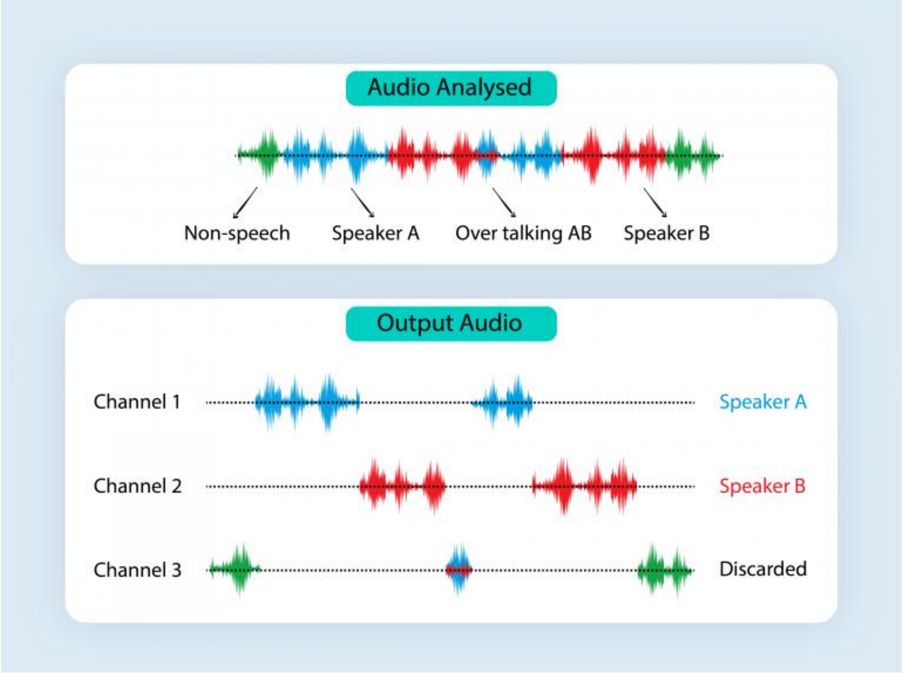
- Replay
- Voice Parody

- Replay attack detection
- Computer generated voice detection (TTS)
- Computer modified voice detection (voice conversion)

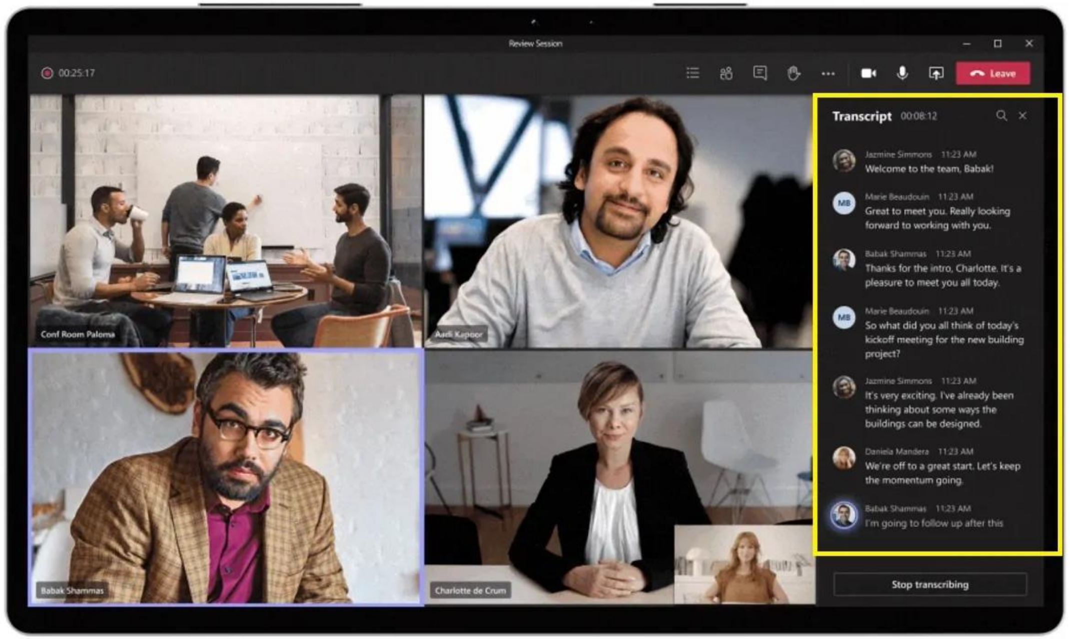
Speaker Diarization

Speaker Diarization is the task of segmenting and co-indexing audio recordings by speaker.

Diarization implies finding speaker boundaries and grouping segments that belong to the same speaker, and, as a by-product, determining the number of distinct speakers.



Applications



Transcript 00:08:12

Jazmine Simmons 11:23 AM
Welcome to the team, Babak!

Marie Beaudouin 11:23 AM
Great to meet you. Really looking forward to working with you.

Babak Shammas 11:23 AM
Thanks for the intro, Charlotte. It's a pleasure to meet you all today.

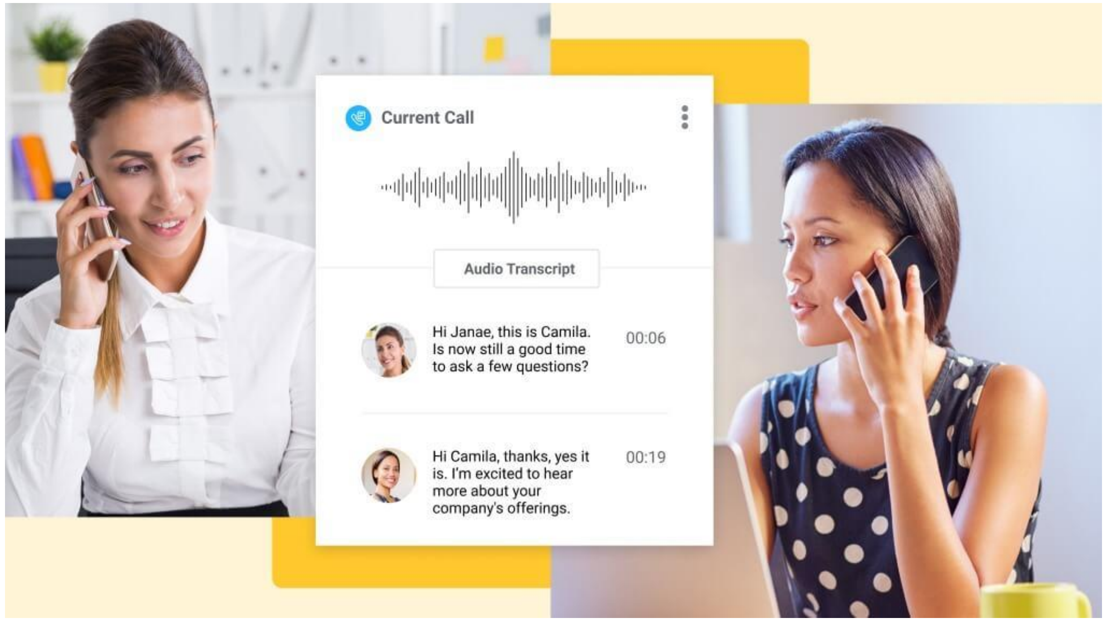
Marie Beaudouin 11:23 AM
So what did you all think of today's kickoff meeting for the new building project?

Jazmine Simmons 11:23 AM
It's very exciting. I've already been thinking about some ways the buildings can be designed.

Daniela Mandera 11:23 AM
We're off to a great start. Let's keep the momentum going.

Babak Shammas 11:23 AM
I'm going to follow up after this

Meeting transcription



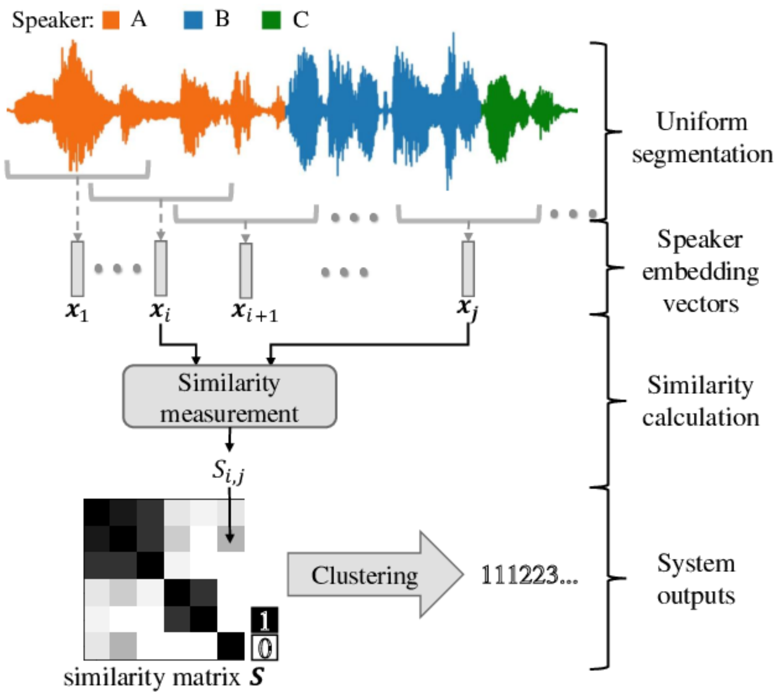
Hi Janae, this is Camila. Is now still a good time to ask a few questions? 00:06

Hi Camila, thanks, yes it is. I'm excited to hear more about your company's offerings. 00:19

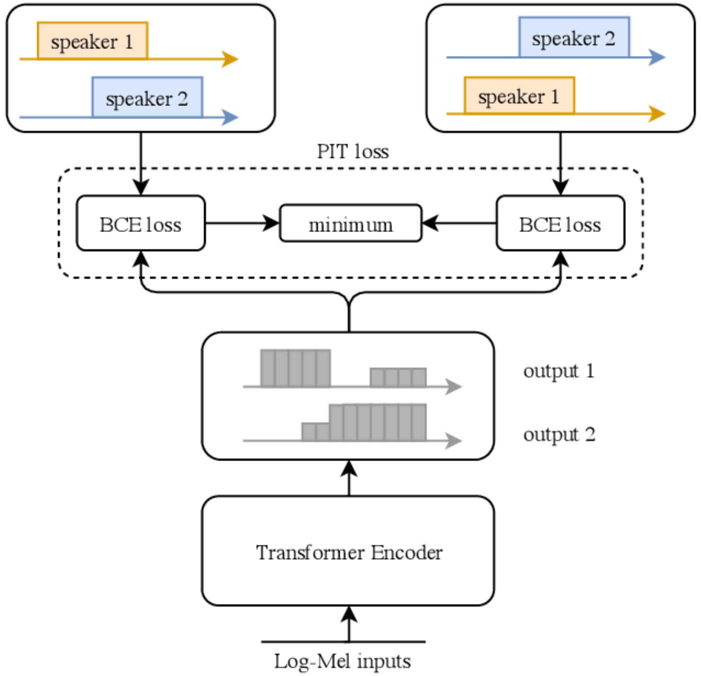
Call transcription

Speaker Diarization

Clustering

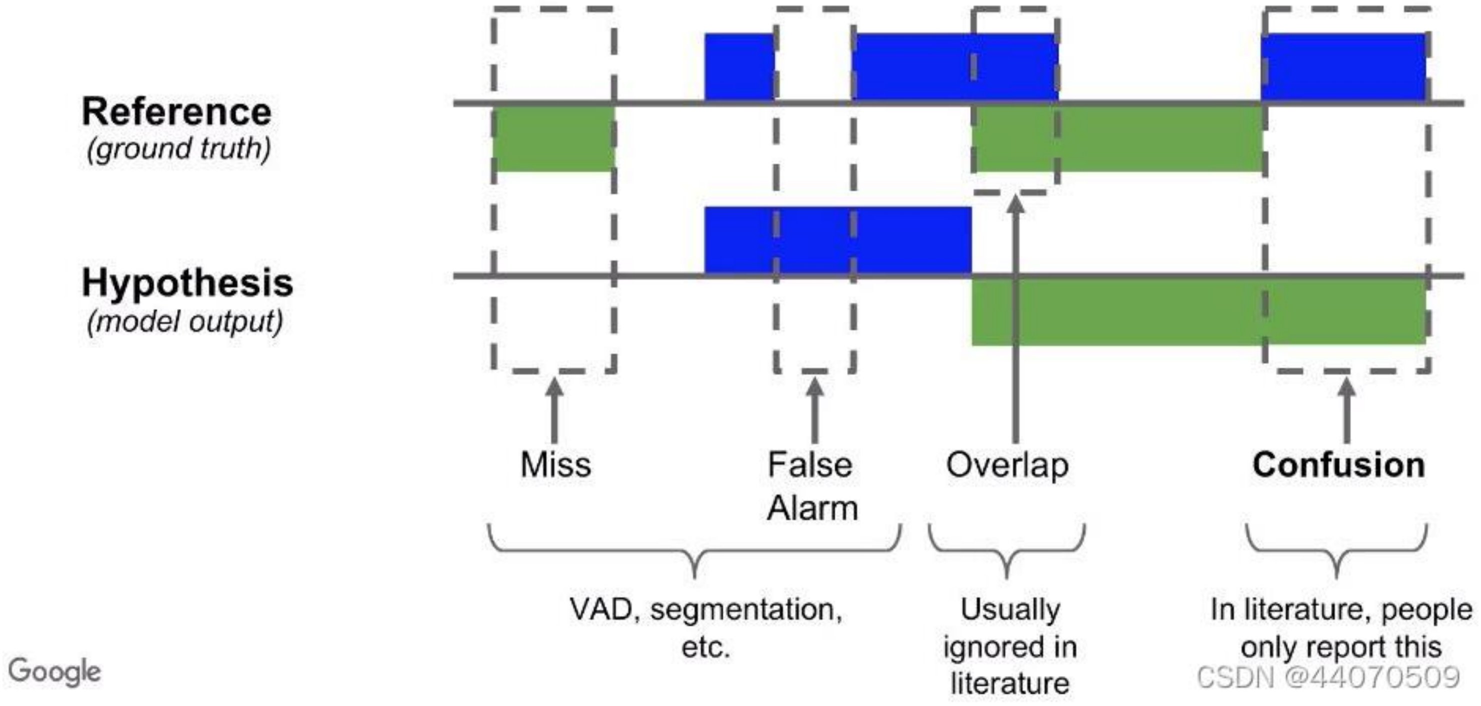


End to end



Metrics

Diarization Error Rate: $$\mathrm{DER} = \frac{T_{FA} + T_{MISS} + T_{SPKR}}{T_{SPEECH}}$$



I. Introduce

II. Speaker Recognition

III. Speaker Classification

Speaker Classification

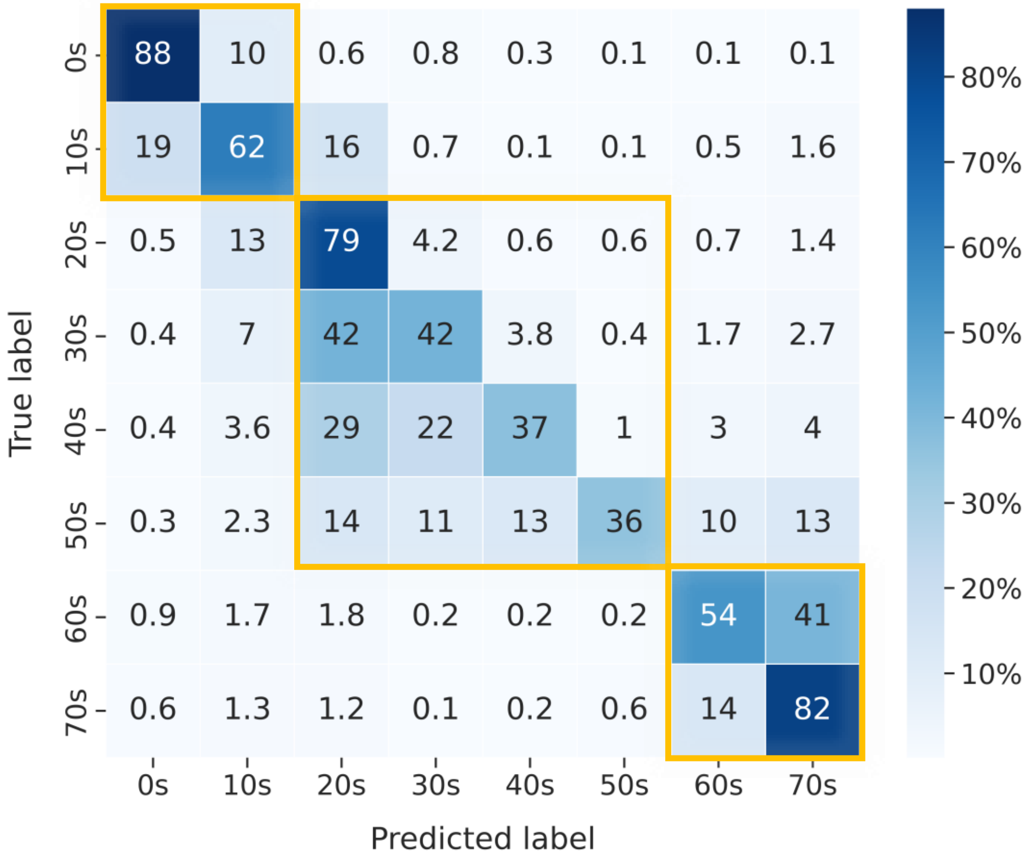
Classify a speaker by:

- Language, accent
- Age
- Gender
- Emotions

Customer benefit :

- Easy UI Language switch
- Gender/Age specific UI
- Enables parental control

Age Estimation



True label \ Predicted label	0s	10s	20s	30s	40s	50s	60s	70s
0s	88	10	0.6	0.8	0.3	0.1	0.1	0.1
10s	19	62	16	0.7	0.1	0.1	0.5	1.6
20s	0.5	13	79	4.2	0.6	0.6	0.7	1.4
30s	0.4	7	42	42	3.8	0.4	1.7	2.7
40s	0.4	3.6	29	22	37	1	3	4
50s	0.3	2.3	14	11	13	36	10	13
60s	0.9	1.7	1.8	0.2	0.2	0.2	54	41
70s	0.6	1.3	1.2	0.1	0.2	0.6	14	82

Age range

- 0-20 years old
- 20-60 years old
- > 60 years old

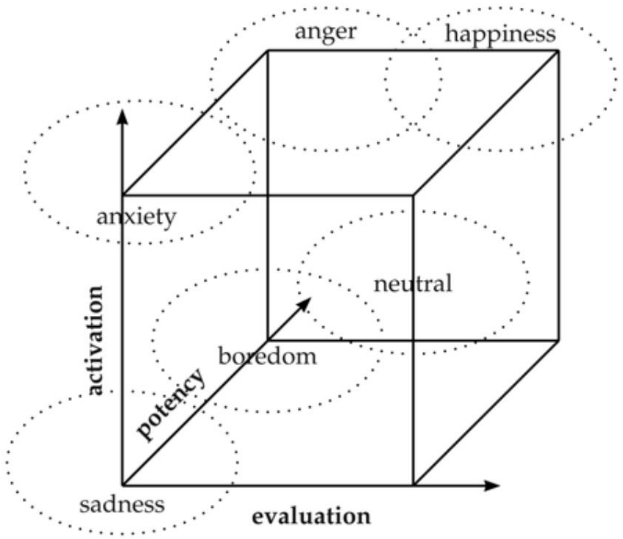
Emotions

Happiness
Sadness
Fear
Anger
Disgust
Surprise

Emotions describe subjective feelings in short periods of time that are related to events, persons, or objects

Approaches:

- Categorical emotion approach
- Dimensional emotion approach



Our Project

Đặc điểm nhận dạng / *Personal indentication:*
Ngày, tháng, năm / *Date, month, year*
CỤC TRƯỞNG CỤC CẢNH SÁT
QUẢN LÝ HÀNH CHÍNH VỀ TRẬT TỰ XÃ HỘI

```
A B C D E 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 A 0 0 0 0 0 0
A B C D < < A B C < A B C <
```

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc
SOCIALIST REPUBLIC OF VIETNAM
Independence - Freedom - Happiness

CĂN CƯỚC CÔNG DÂN
Citizen Identity Card

Số / *No.:*
Họ và tên / *Full name:*
Ngày sinh / *Date of birth:*
Giới tính / *Sex:* Quốc tịch / *Nationality:*
Quê quán / *Place of origin:*
Có giá trị đến: *Date of expiry:*
Nơi thường trú / *Place of residence:*

FWD
INSURANCE

VINBIGDATA

Thank you

Tầng 9, Century Tower, Times City, 458 Minh Khai, Phường
Vĩnh Tuy, Quận Hai Bà Trưng, Hà Nội. 📱

info@vinbigdata.org ✉

product.vinbigdata.org 🔗