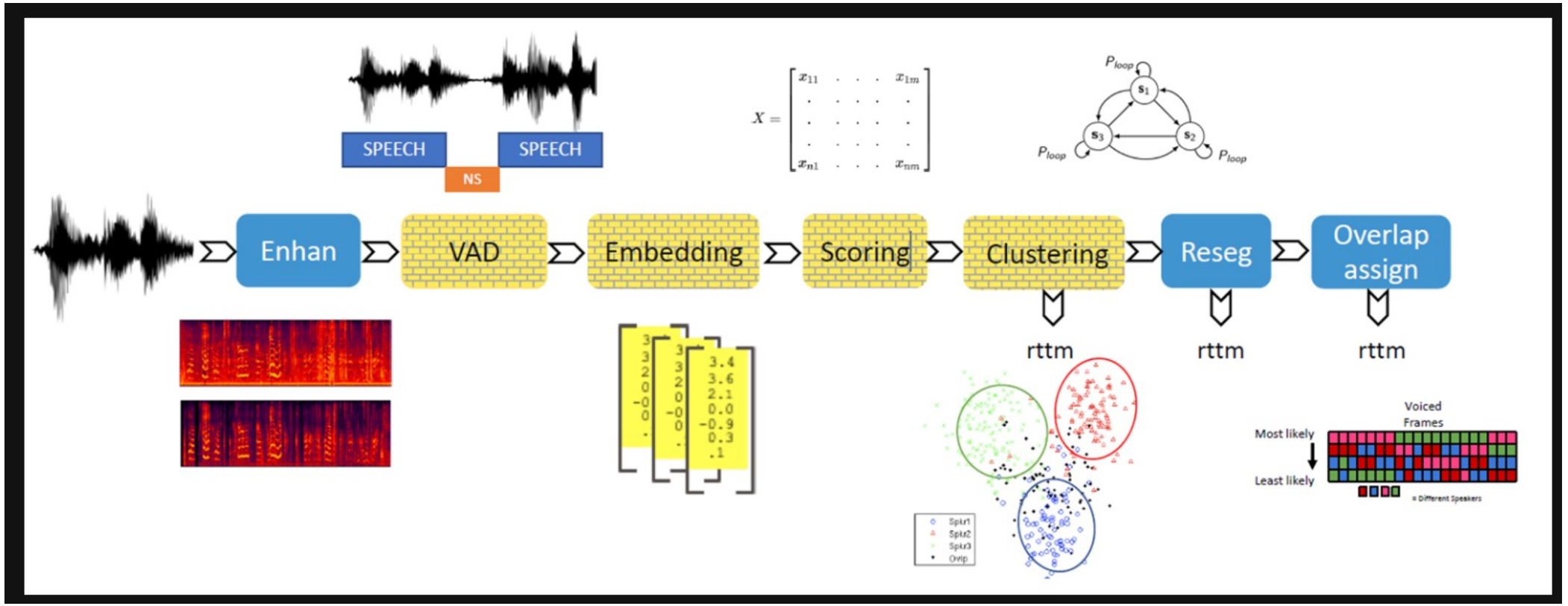
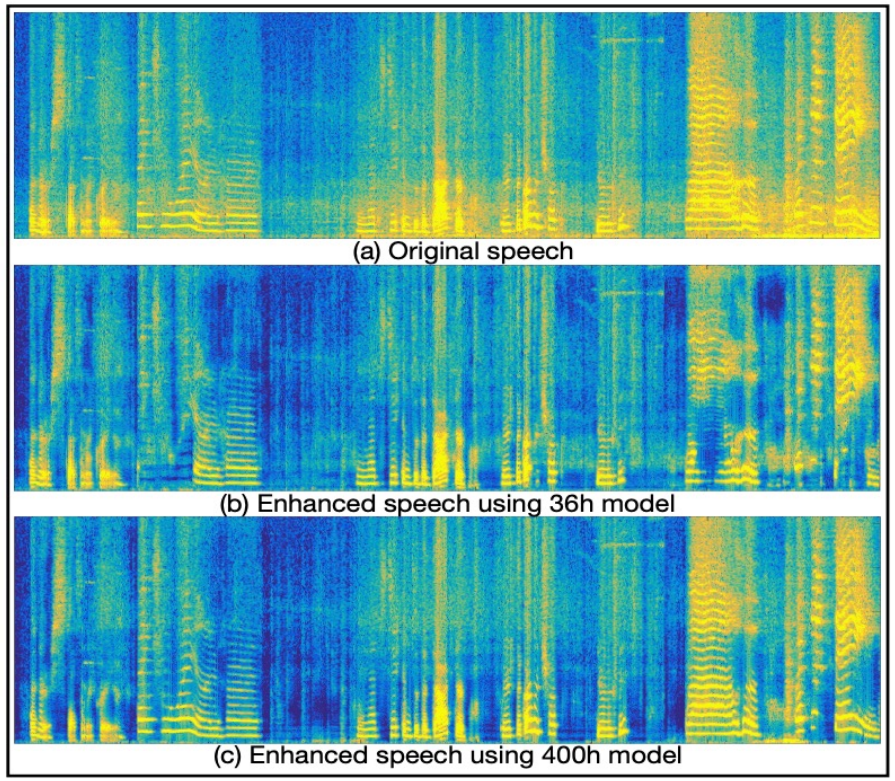
ANALYSIS OF THE BUT DIARIZATION SYSTEM FOR VOXCONVERSE CHALLENGE

Overview


SPEECH
NS
SPEECH
Enhan
VAD
Embedding
Scoring
Clustering
Reseg
Overlap assign
rttm
rttm
rttm
Most likely
Least likely
Voiced Frames

1. Preprocessing

LSTM을 통한 Enhancement



Background noise를 최대한 지우면서
speaker specific information은 보존해야한다.

Trade-off 관계

해당 모델은 pre trained LSTM을 이용

VAD(Voice Activity Detection)

해당 구간이 Speech? Non-Speech?

- energy-based VAD
  Voice Frame은 energy가silent보다 많을 것이라는 가정으로 출발
- deep neural network (DNN) based system
  25ms씩 끊고 10ms씩 움직여서 40차원의 Filter bank를 만듬
  ➔이후 앞 frame 5개, 뒤 frame5개를 concat해서 440 차원을 만듬
  ➔Linear layer를 통과시켜 0과1로 classification
- automatic speech recognition (ASR) based system
  Kaldi toolkit을 사용하여 음소 단위 label 생성

- Energy, DNN 모델에는 median filter를 사용

VAD(Voice Activity Detection)

해당 구간이 Speech? Non-Speech?

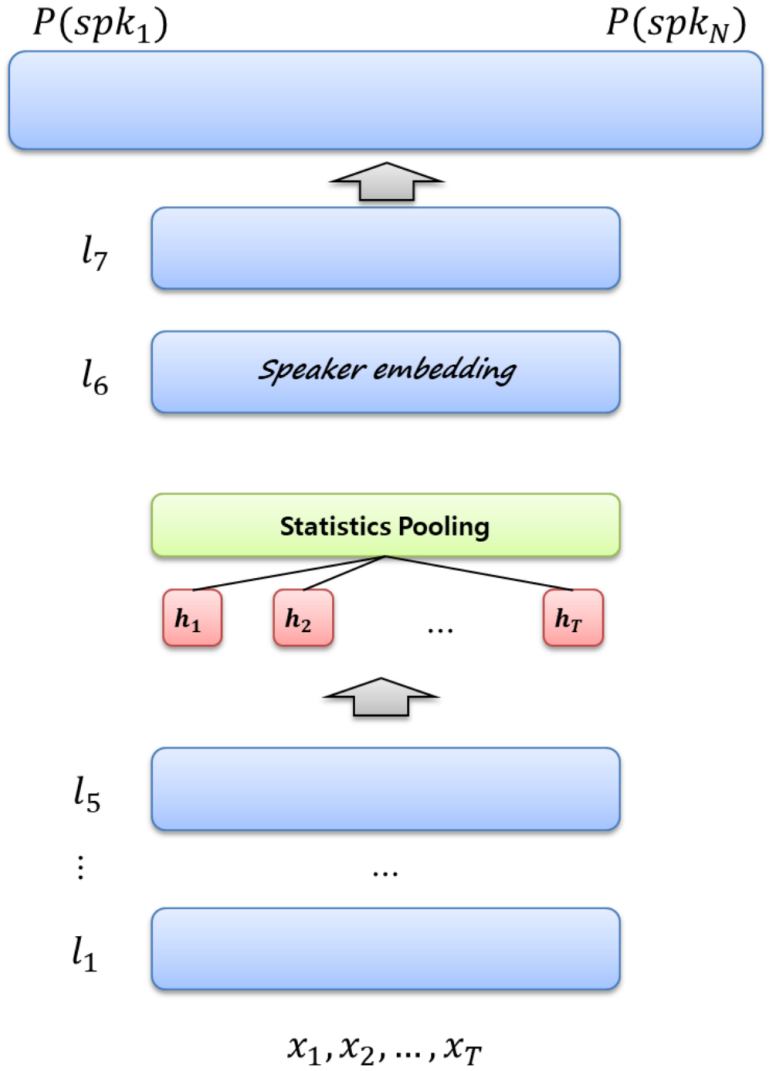
- VAD 시스템 분석 결과 대부분이 짧은 조용한 구간(short periods of silence)을 기준으로 speech segmentation을 진행(말을 하고 있어도 잠깐 짧은 조용한 구간이 나오면 segment)

- 따라서 오류를 줄이기 위해 특정 length보다 짧은 silence는 speech로 label (hyperparameter)

- 이후 위에서 말한 vad system을 ensemble함(voting) -> 대회라서 가능

Speaker embeddings

Speech 구간에서 speaker의 feature extraction -> x-vector



TDNN(Time Delay Neural Network) = 1-D Conv 를 이용해서 N - speaker classification을 수행

이후 Ln번째 layer의 값을 embedding vector로 취함

저자들은 152-layer resnet을 이용하여 5994 speaker Classification을 수행 후 256dim의 vector를 extract.

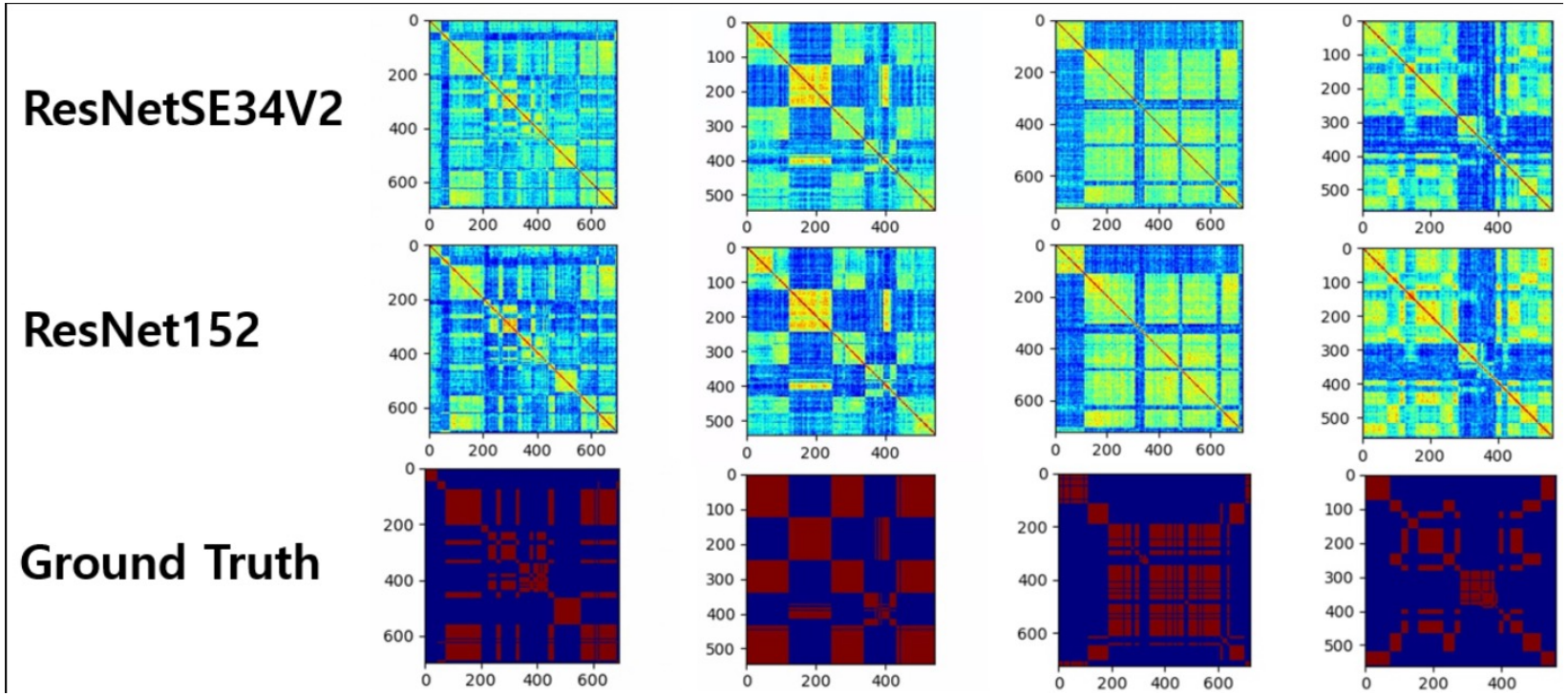
Data는 2초의 chunk로 나눠 noise를 통해 augment하고 25ms의 커널크기와 10ms의 stride를 가지고 64dim의 filter bank를 만듦 -> filter bank라는게 filter 크기로 생각해도 될거 같기도..?

Loss:

$$L_{ns} = \frac{1}{N}\sum_i -\log \frac{e^{s\cos(\theta_{y_i,i})}}{\sum_j e^{s\cos(\theta_{j,i})}}.$$

Initial clustering

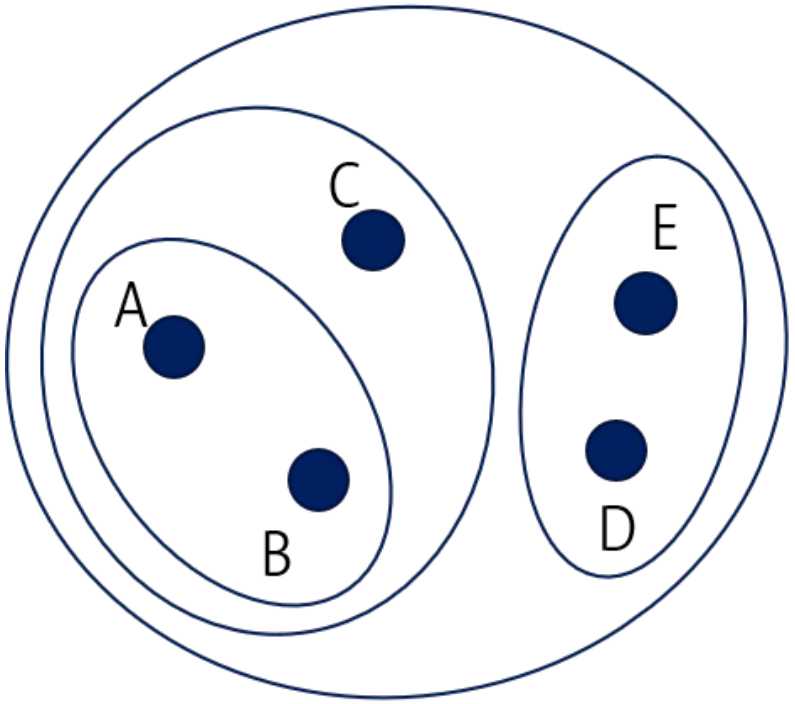
1. probabilistic linear discriminant analysis (PLDA) 를 이용해 x-vector간의 score 계산(얼마나 유사한지) -> cosine similarity도 가능



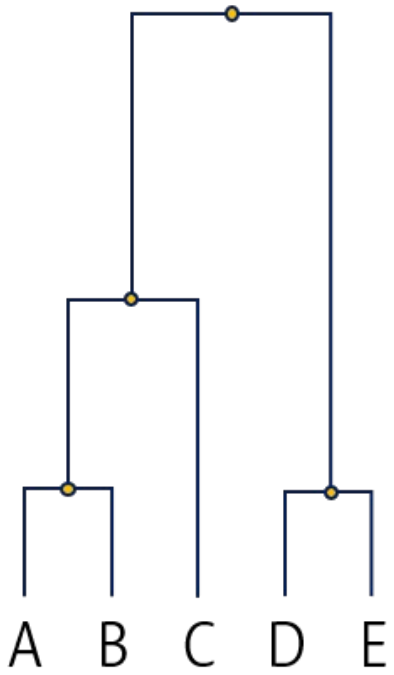
Initial clustering

2. agglomerative hierarchical clustering(AHC)를 이용해 clustering

Nested clusters



Dendrogram



Initial clustering

3. BUT AHC는 cluster 갯수가 하이퍼 마라미터로 이에 민감.

SO Bayesian HMM를 사용

HMM의 state는 represent speaker로 설정하고

Transition과 state distribution은 앞서 train한 PLDA 모델에서 가져옴

무슨말인지 모르겠습니다..

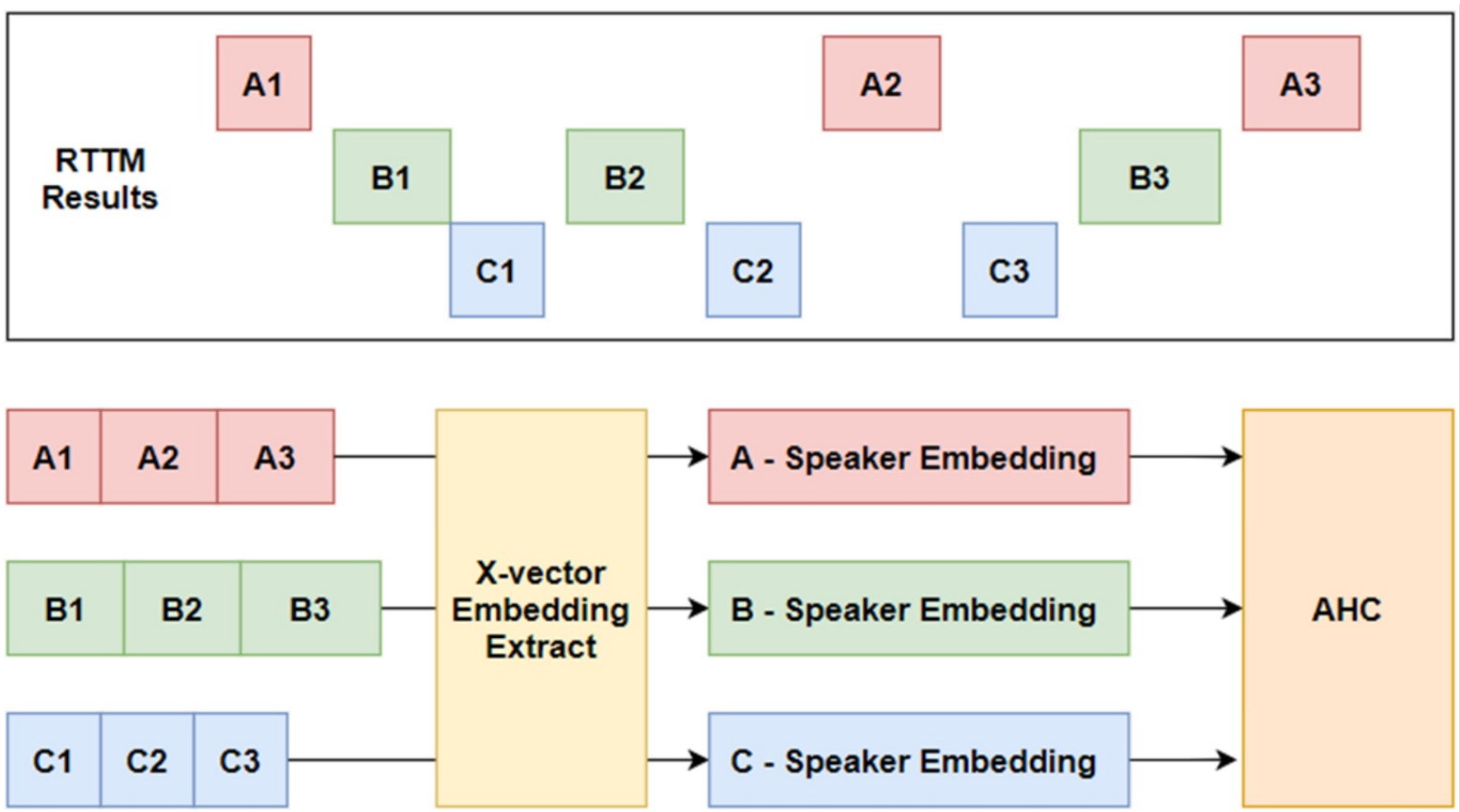
reclustering

1. 초기에 segment가 short time으로 segmentation됐다.

2. 만약 임베딩이 더 긴 segment에서 나온다면 더 robust하고 clustering이 잘될 것이다.

3. 앞서 나온 speaker를 바탕으로 same speaker끼리 concat해서 x-vector Embedding 다시 실시

4. 위에서 나온 방식대로 scoring 후 같은 화자라고 판단되면 join


RTTM Results
A1
B1
C1
A2
B2
C2
A3
B3
C3
A1
A2
A3
B1
B2
B3
C1
C2
C3
X-vector Embedding Extract
A - Speaker Embedding
B - Speaker Embedding
C - Speaker Embedding
AHC

Overlapped speech handling

- VAD는 segment당 speech가 있냐 없냐만 판단.
- 말이 겹치는 부분은 VAD처럼 overlap만 감지하는 모델을 만들어서 찾아냄.
- VAD와 비슷하게 학습

Overlapped speech handling

Overlap된 부분을 VB-HMM을 통해 similarity score를 구한 후 두번째로 score가 높은 speaker가 Overlap되었다고 판단