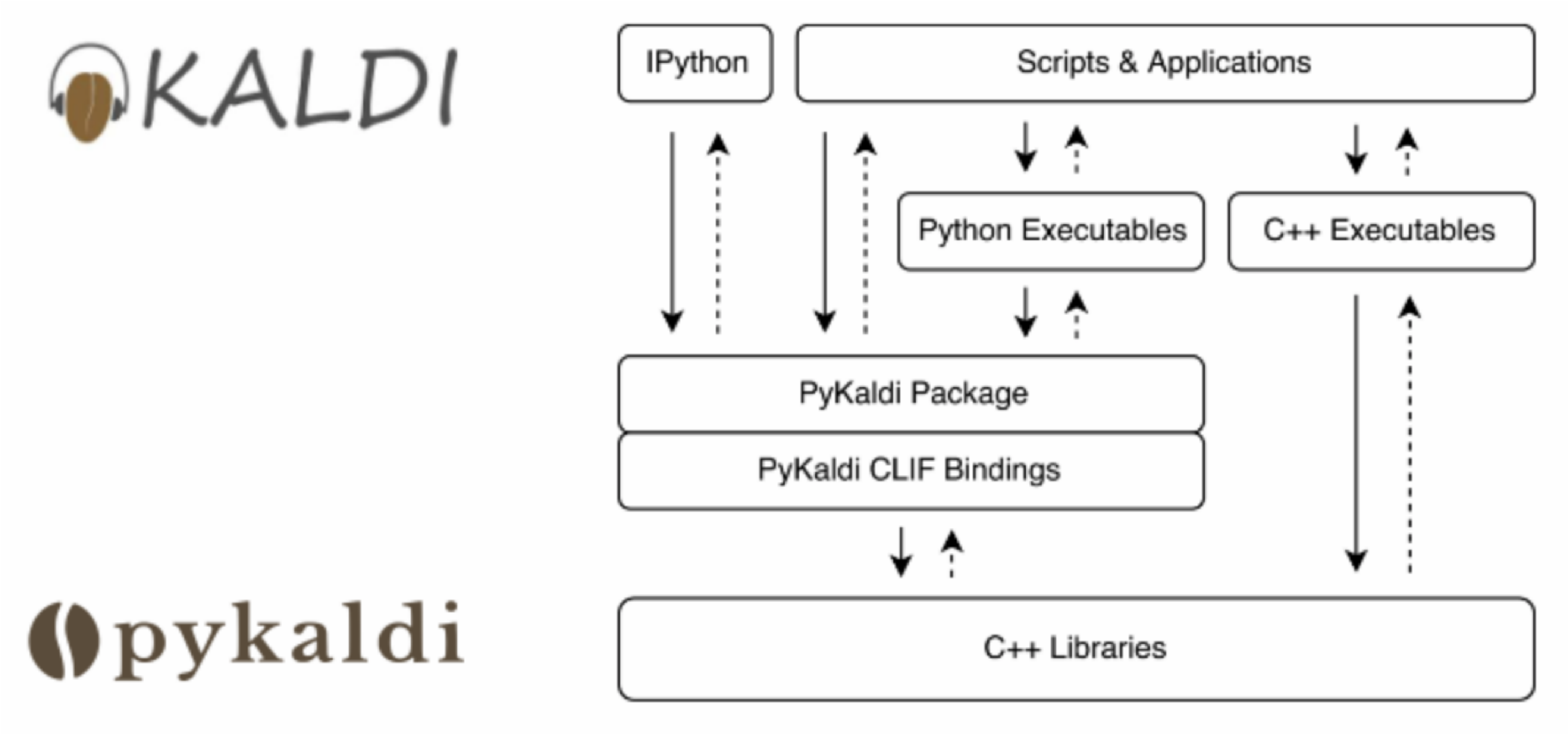
Speech Recognition using PyKaldi

Yuan-Fu Liao

National Taipei University of Technology

PyKaldi: A Python wrapper for Kaldi



IPython

Scripts & Applications

Python Executables

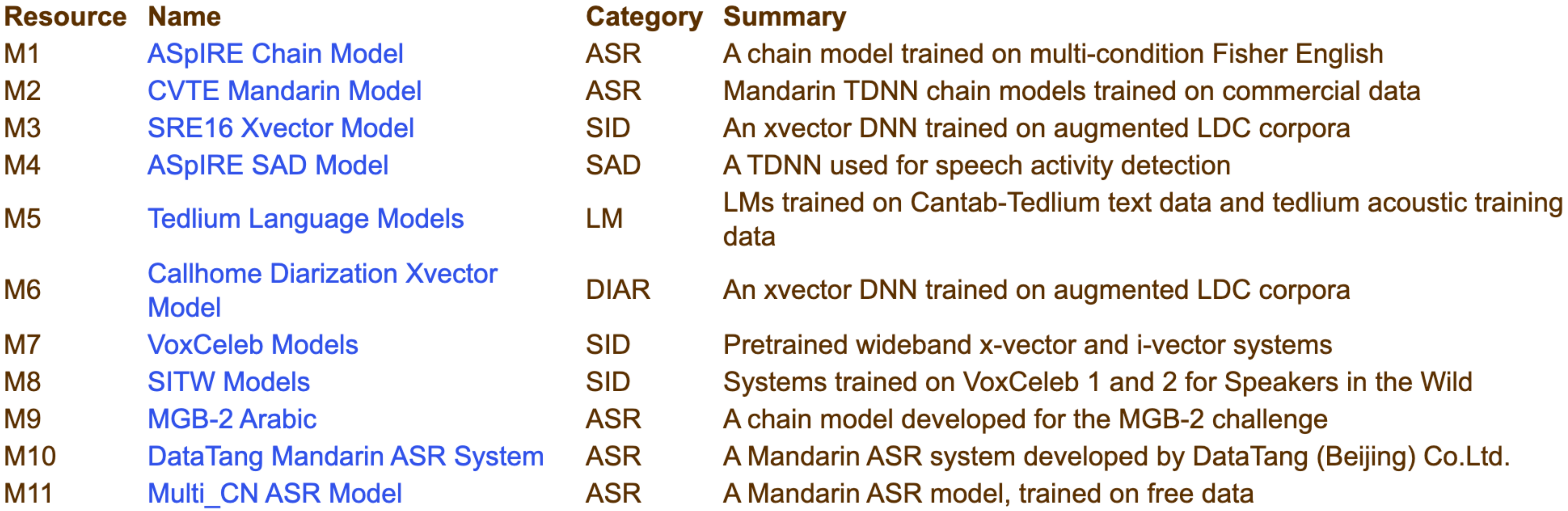
C++ Executables

PyKaldi Package

PyKaldi CLIF Bindings

C++ Libraries

Pre-Trained Models



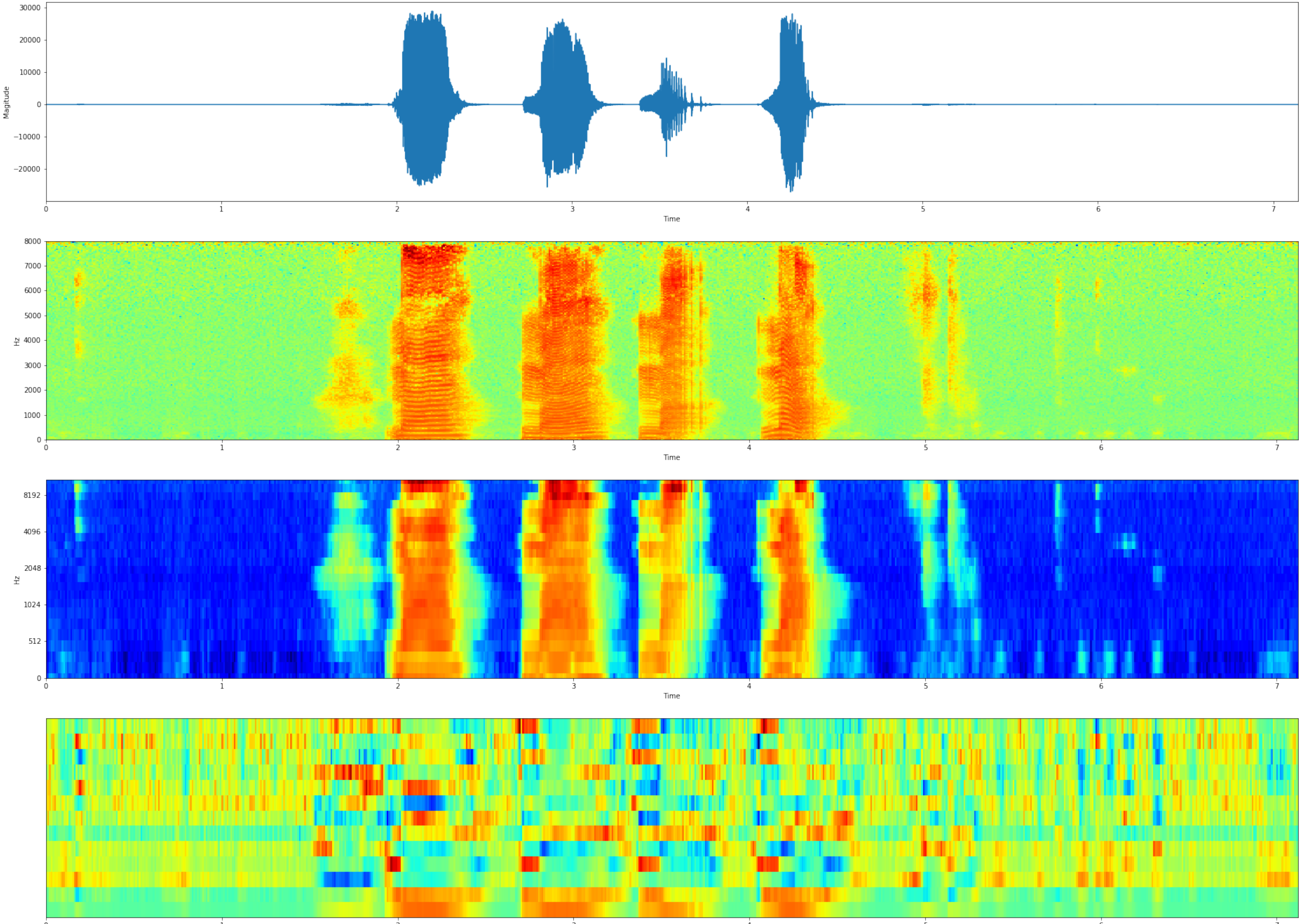
Resource	Name	Category	Summary
M1	ASpIRE Chain Model	ASR	A chain model trained on multi-condition Fisher English
M2	CVTE Mandarin Model	ASR	Mandarin TDNN chain models trained on commercial data
M3	SRE16 Xvector Model	SID	An xvector DNN trained on augmented LDC corpora
M4	ASpIRE SAD Model	SAD	A TDNN used for speech activity detection
M5	Tedlium Language Models	LM	LMs trained on Cantab-Tedlium text data and tedlium acoustic training data
M6	Callhome Diarization Xvector Model	DIAR	An xvector DNN trained on augmented LDC corpora
M7	VoxCeleb Models	SID	Pretrained wideband x-vector and i-vector systems
M8	SITW Models	SID	Systems trained on VoxCeleb 1 and 2 for Speakers in the Wild
M9	MGB-2 Arabic	ASR	A chain model developed for the MGB-2 challenge
M10	DataTang Mandarin ASR System	ASR	A Mandarin ASR system developed by DataTang (Beijing) Co.Ltd.
M11	Multi_CN ASR Model	ASR	A Mandarin ASR model, trained on free data

Installation

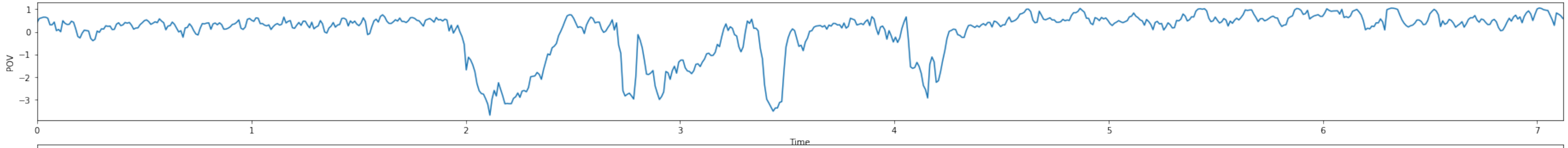
- Conda
  - conda install -c pykaldi pykaldi
  - ……
- From Source
  - git clone https://github.com/pykaldi/pykaldi.git
  - Tools
  - Setup
  - Test
  - ……

Feature Extraction

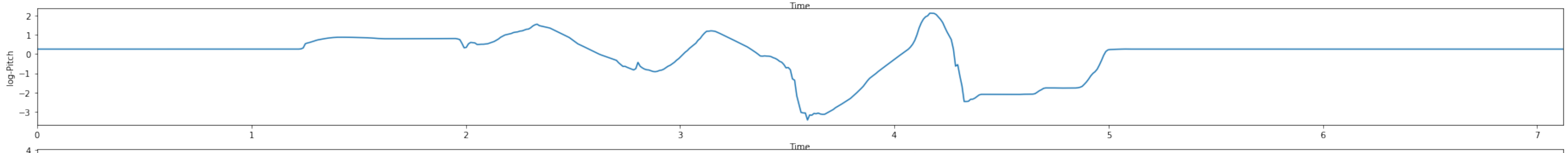
- Waveform
- Spectrogram
- MelFBank
- MFCCs



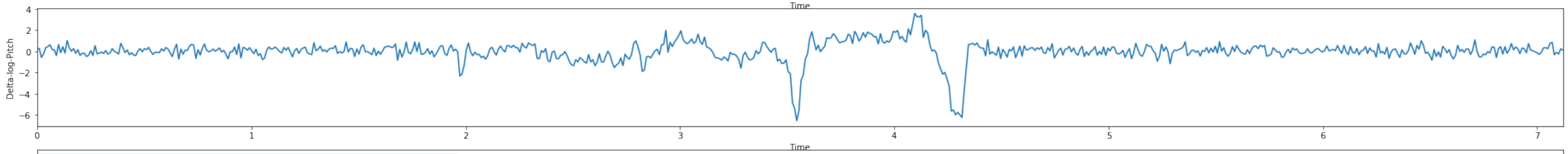
- POV



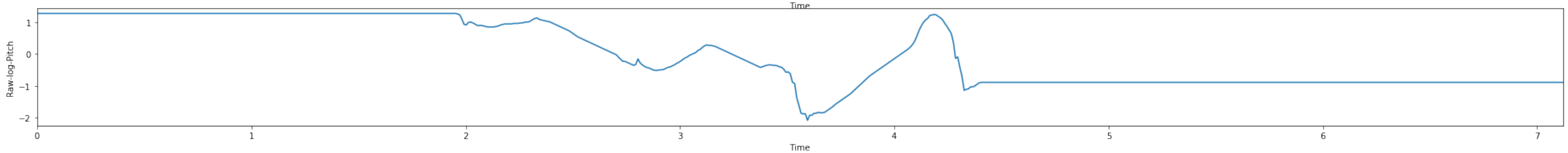
- Pitch



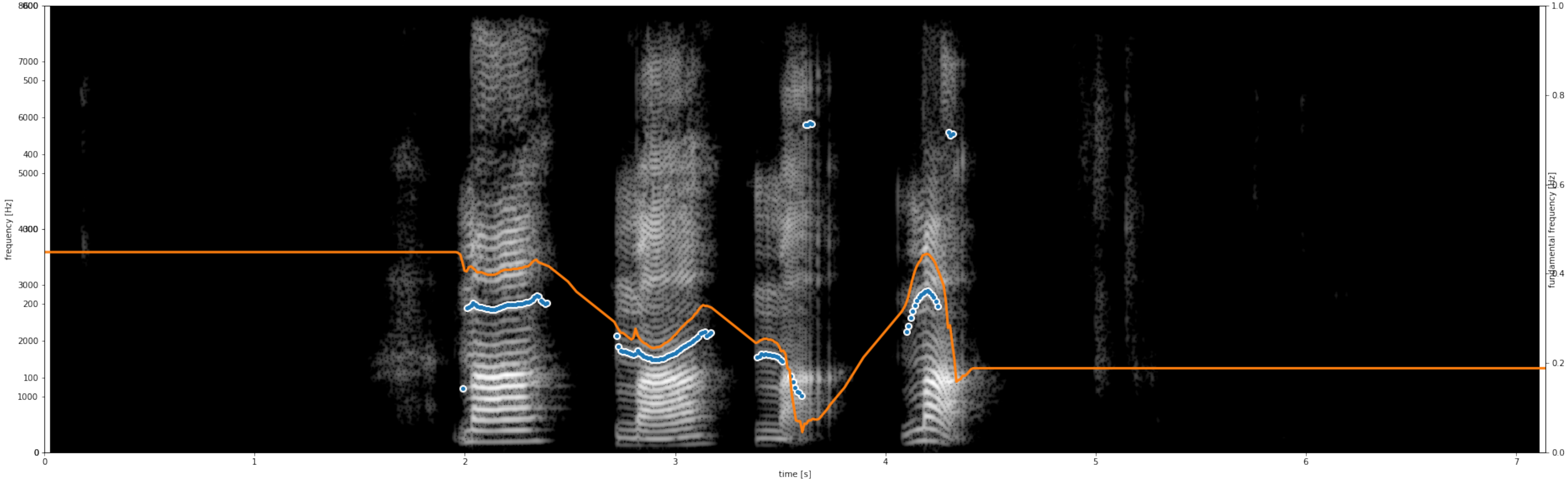
- Delta-Pitch



- Raw-Pitch



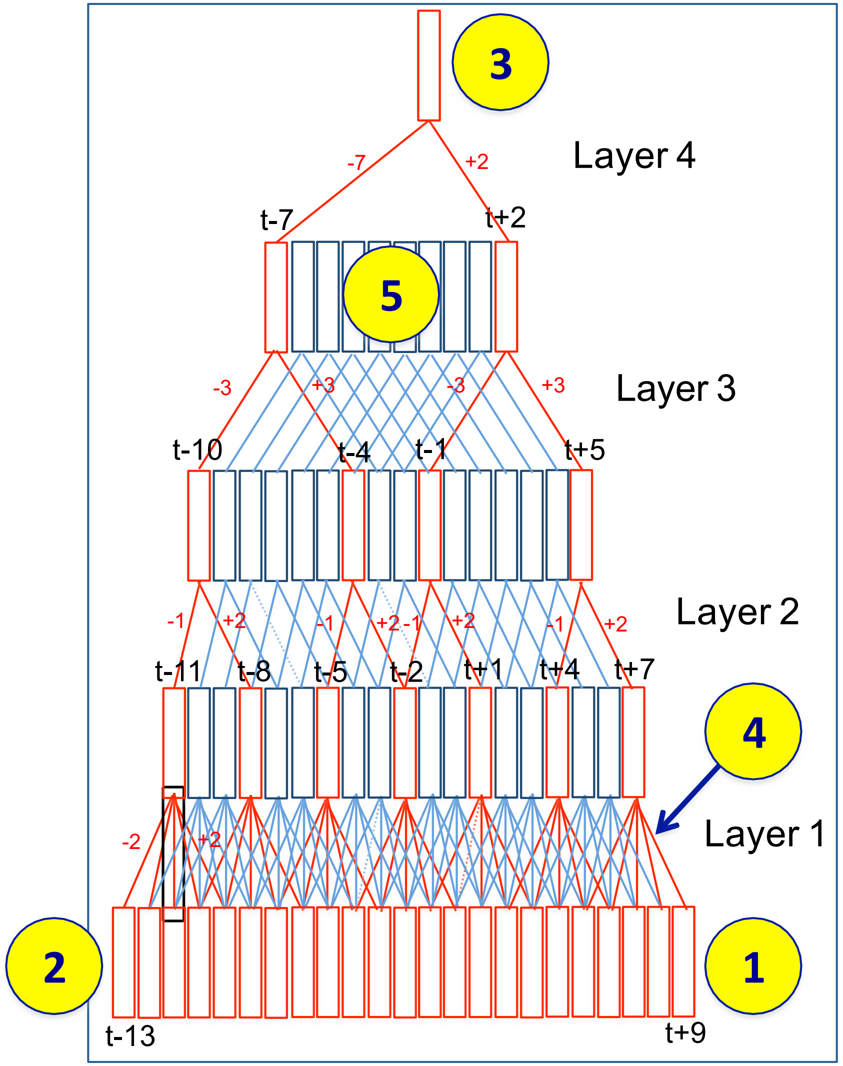
- Pitch: Kaldi vs. Praat



Recipe: ASpIRE – English Speech Recognizer

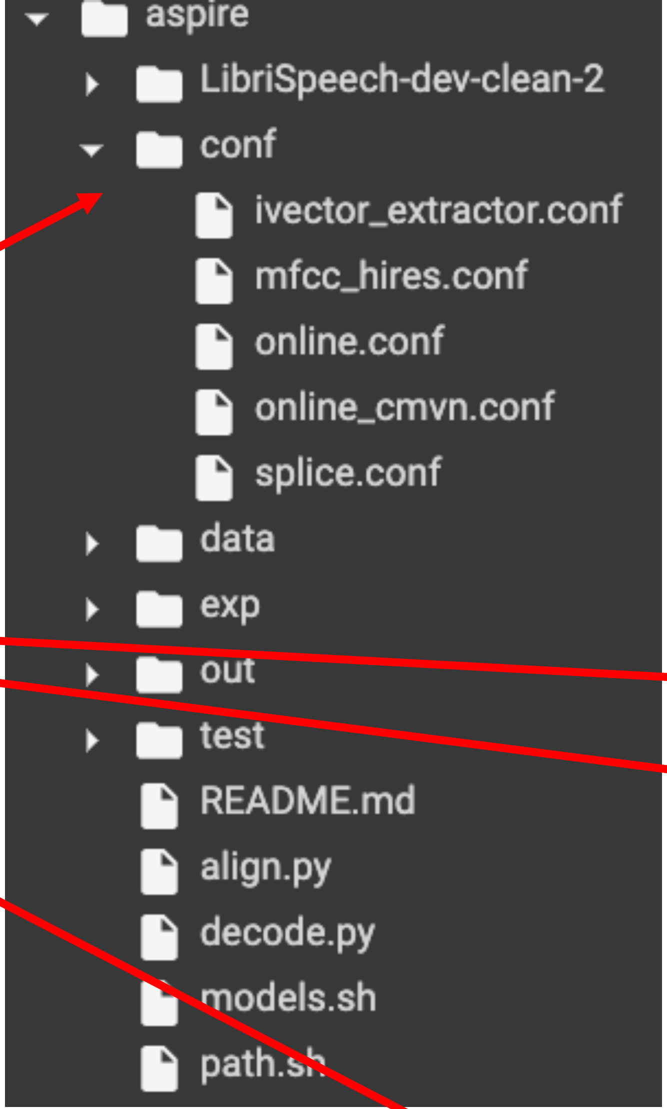
- Reverberant Environments Challenge
  - Time delay neural networks (TDNN, chain model)
  - Data augmentation with simulated reverberations
  - i-vector based speaker & environment adaptation

Voice Enabled Smart-Home

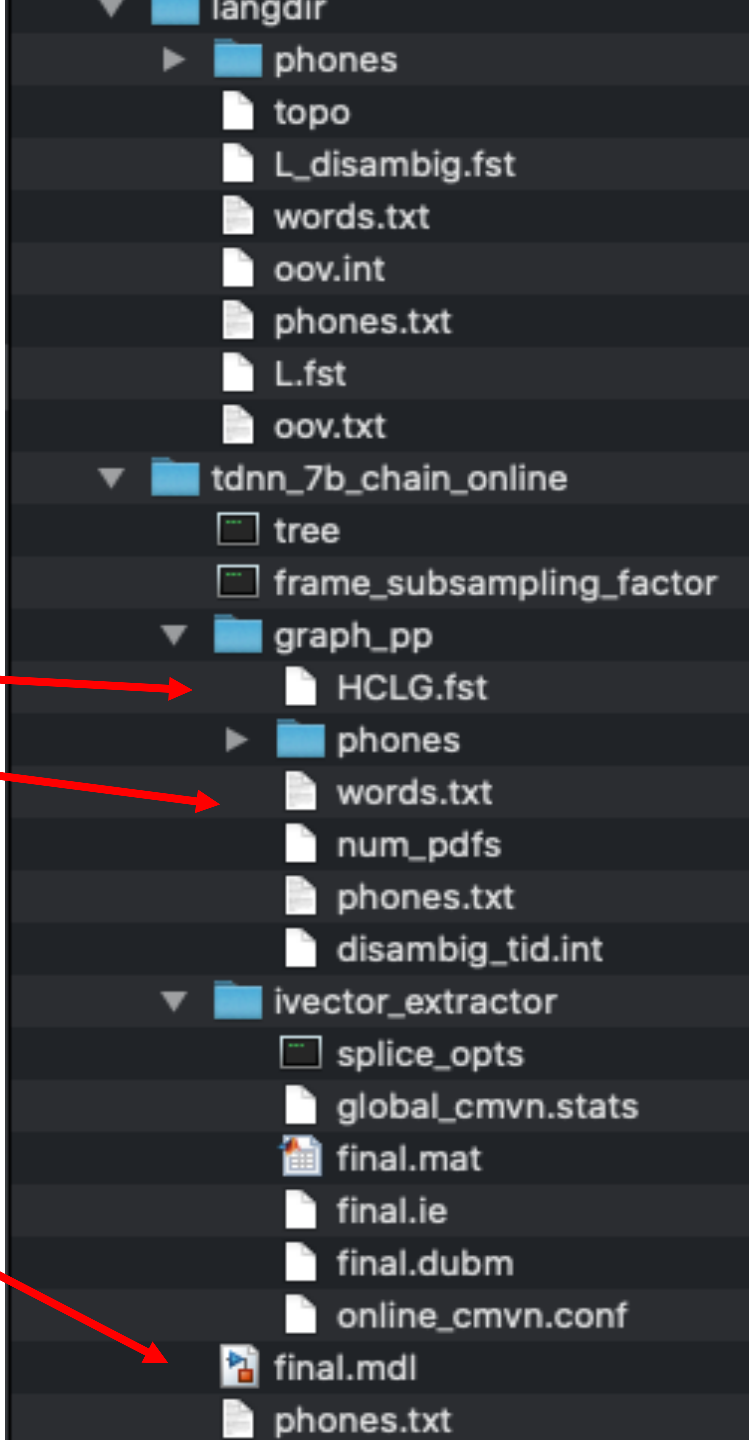


Pre-Trained Model

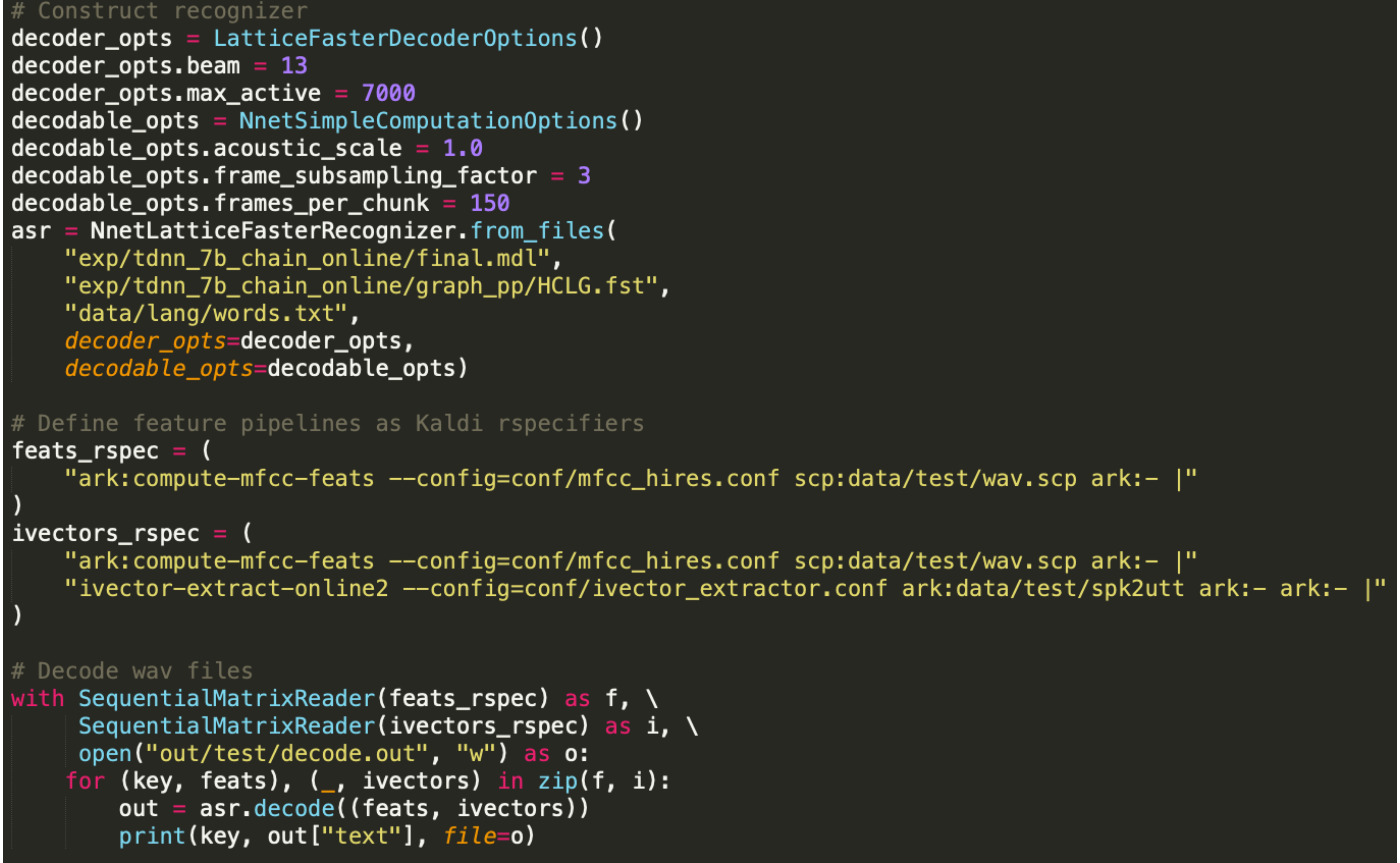
- NER-Trs-Vol1
  - Config
  - Model



```
aspire
    LibriSpeech-dev-clean-2
    conf
        ivector_extractor.conf
        mfcc_hires.conf
        online.conf
        online_cmvn.conf
        splice.conf
    data
    exp
    out
    test
    README.md
    align.py
    decode.py
    models.sh
    path.sh
```



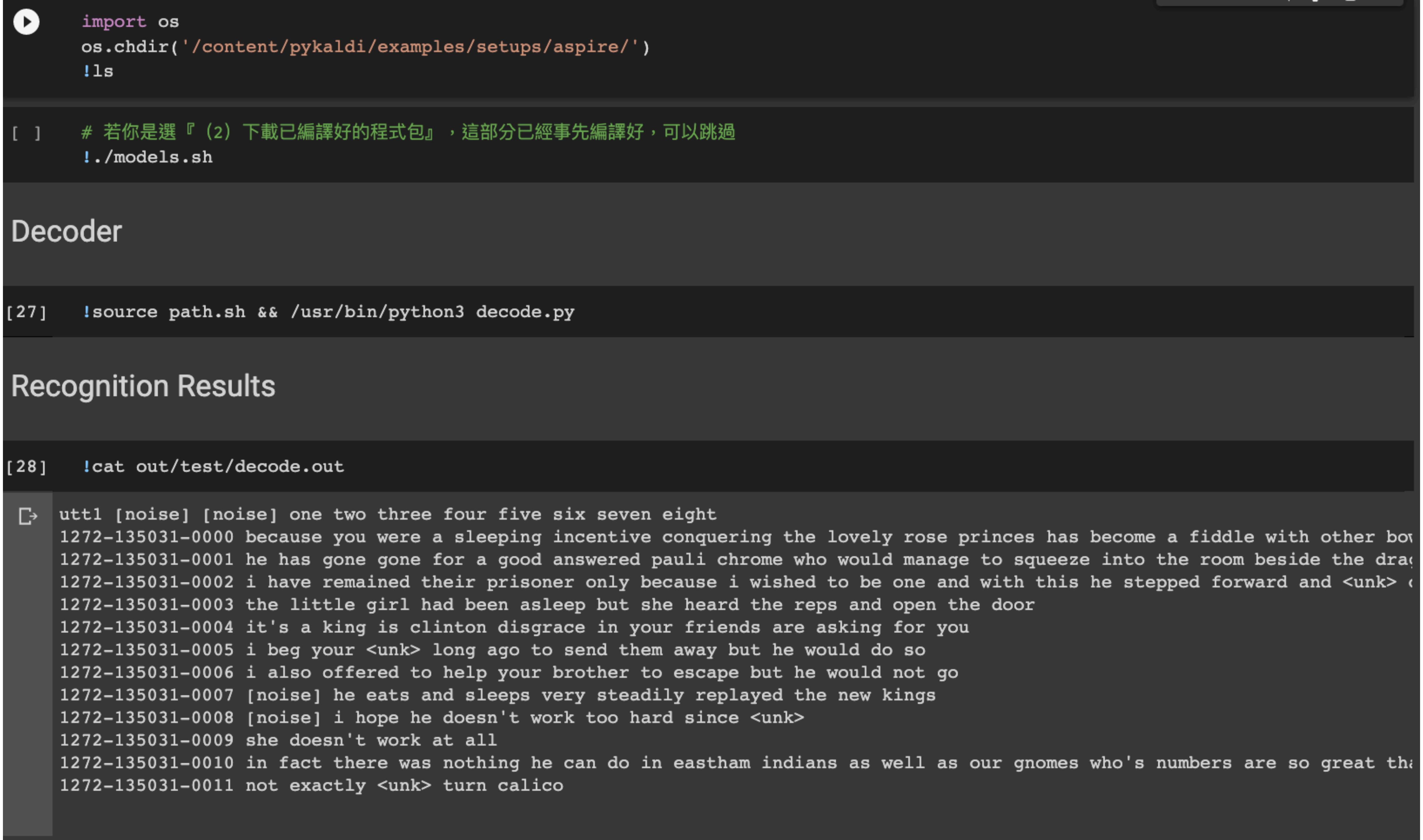
```
langdir
    phones
    topo
    L_disambig.fst
    words.txt
    oov.int
    phones.txt
    L.fst
    oov.txt
tdnn_7b_chain_online
    tree
    frame_subsampling_factor
    graph_pp
        HCLG.fst
        phones
        words.txt
        num_pdfs
        phones.txt
        disambig_tid.int
    ivector_extractor
        splice_opts
        global_cmvn.stats
        final.mat
        final.ie
        final.dubm
        online_cmvn.conf
    final.mdl
    phones.txt
```



```
# Construct recognizer
decoder_opts = LatticeFasterDecoderOptions()
decoder_opts.beam = 13
decoder_opts.max_active = 7000
decodable_opts = NnetSimpleComputationOptions()
decodable_opts.acoustic_scale = 1.0
decodable_opts.frame_subsampling_factor = 3
decodable_opts.frames_per_chunk = 150
asr = NnetLatticeFasterRecognizer.from_files(
    "exp/tdnn_7b_chain_online/final.mdl",
    "exp/tdnn_7b_chain_online/graph_pp/HCLG.fst",
    "data/lang/words.txt",
    decoder_opts=decoder_opts,
    decodable_opts=decodable_opts)

# Define feature pipelines as Kaldi rspecifiers
feats_rspec = (
    "ark:compute-mfcc-feats --config=conf/mfcc_hires.conf scp:data/test/wav.scp ark:- |"
)
ivectors_rspec = (
    "ark:compute-mfcc-feats --config=conf/mfcc_hires.conf scp:data/test/wav.scp ark:- |"
    "ivector-extract-online2 --config=conf/ivector_extractor.conf ark:data/test/spk2utt ark:- ark:- |"
)

# Decode wav files
with SequentialMatrixReader(feats_rspec) as f, \
     SequentialMatrixReader(ivectors_rspec) as i, \
     open("out/test/decode.out", "w") as o:
    for (key, feats), (_, ivectors) in zip(f, i):
        out = asr.decode((feats, ivectors))
        print(key, out["text"], file=o)
```



```
import os
os.chdir('/content/pykaldi/examples/setups/aspire/')
!ls
```

[ ]

```
# 若你是選『（2）下載已編譯好的程式包』，這部分已經事先編譯好，可以跳過
!./models.sh
```

Decoder

[27]

```
!source path.sh && /usr/bin/python3 decode.py
```

Recognition Results

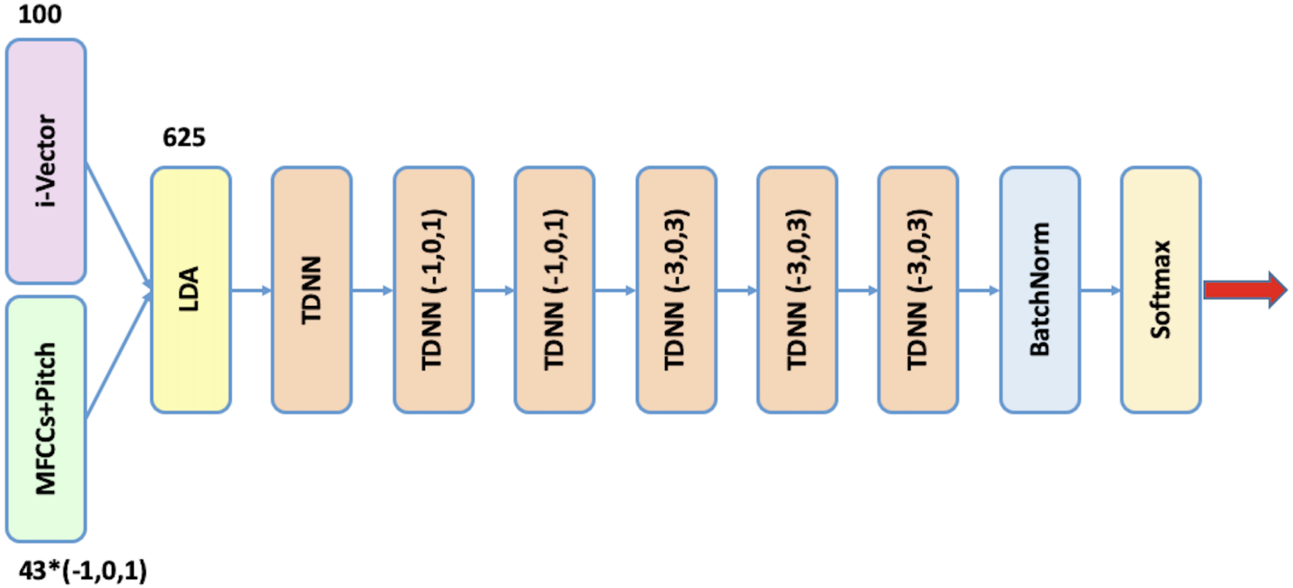
[28]

```
!cat out/test/decode.out
```

```
utt1 [noise] [noise] one two three four five six seven eight
1272-135031-0000 because you were a sleeping incentive conquering the lovely rose princes has become a fiddle with other bo
1272-135031-0001 he has gone gone for a good answered pauli chrome who would manage to squeeze into the room beside the dra
1272-135031-0002 i have remained their prisoner only because i wished to be one and with this he stepped forward and <unk>
1272-135031-0003 the little girl had been asleep but she heard the reps and open the door
1272-135031-0004 it's a king is clinton disgrace in your friends are asking for you
1272-135031-0005 i beg your <unk> long ago to send them away but he would do so
1272-135031-0006 i also offered to help your brother to escape but he would not go
1272-135031-0007 [noise] he eats and sleeps very steadily replayed the new kings
1272-135031-0008 [noise] i hope he doesn't work too hard since <unk>
1272-135031-0009 she doesn't work at all
1272-135031-0010 in fact there was nothing he can do in eastham indians as well as our gnomes who's numbers are so great th
1272-135031-0011 not exactly <unk> turn calico
```

Recipe: NER-Trs-Vol1 – Mandarin Speech Recognizer

- Broadcast Radio Shows



100

i-Vector

MFCCs+Pitch

43*(-1,0,1)

625

LDA

TDNN

TDNN (-1,0,1)

TDNN (-1,0,1)

TDNN (-3,0,3)

TDNN (-3,0,3)

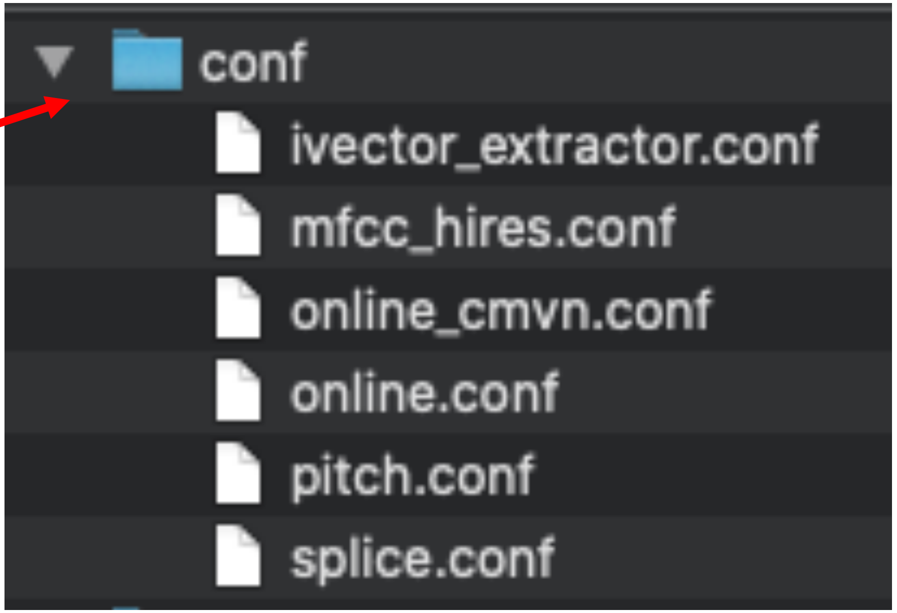
TDNN (-3,0,3)

BatchNorm

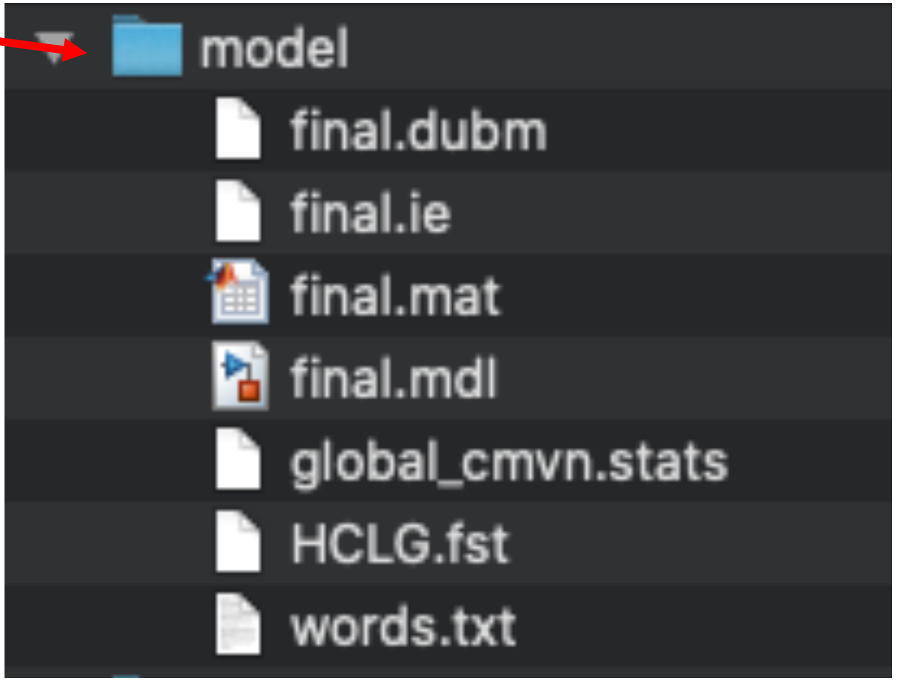
Softmax

Pre-Trained Model

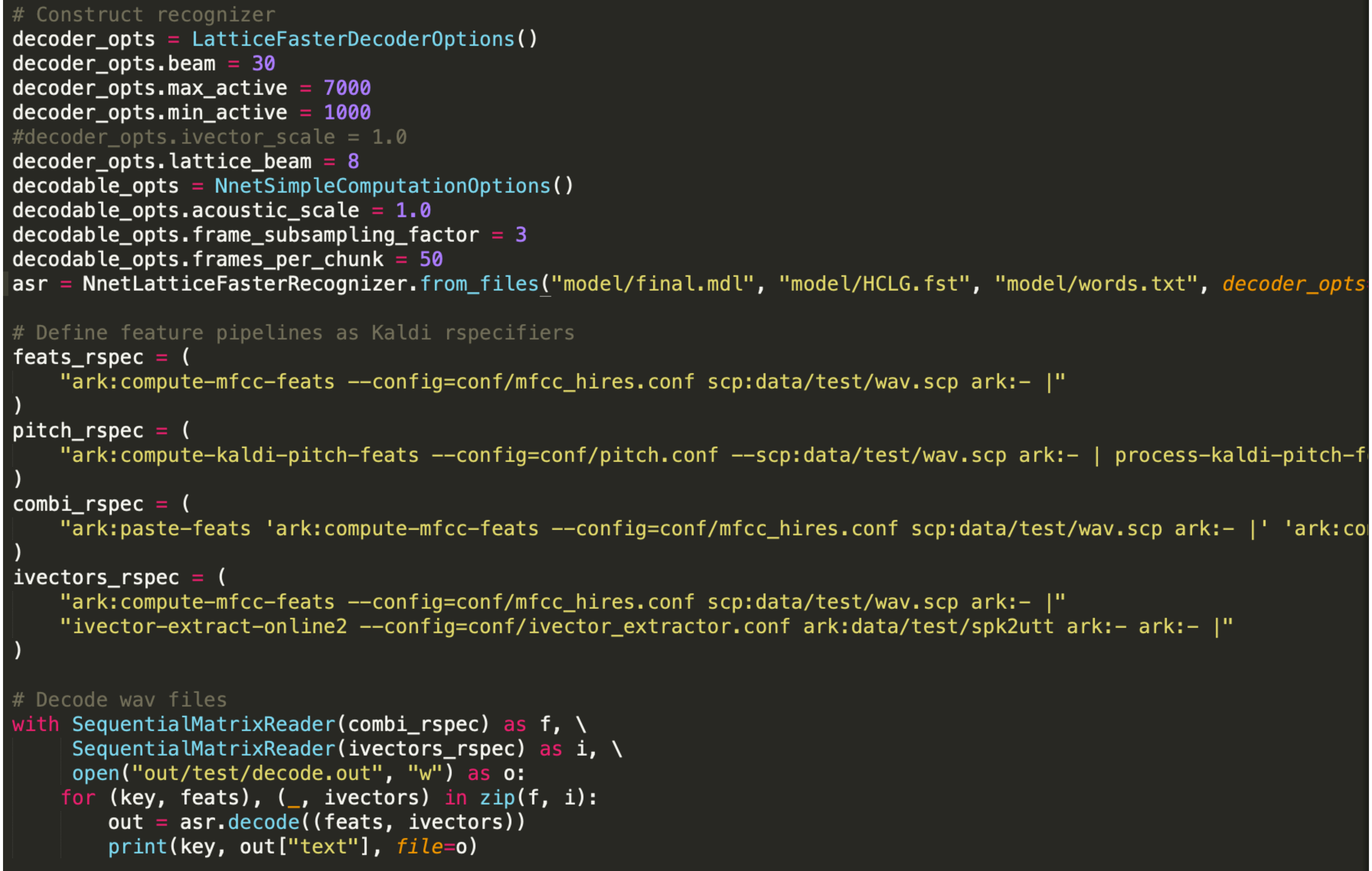
- NER-Trs-Vol1
  - Config
  - Model



- conf
  - ivector_extractor.conf
  - mfcc_hires.conf
  - online_cmvn.conf
  - online.conf
  - pitch.conf
  - splice.conf



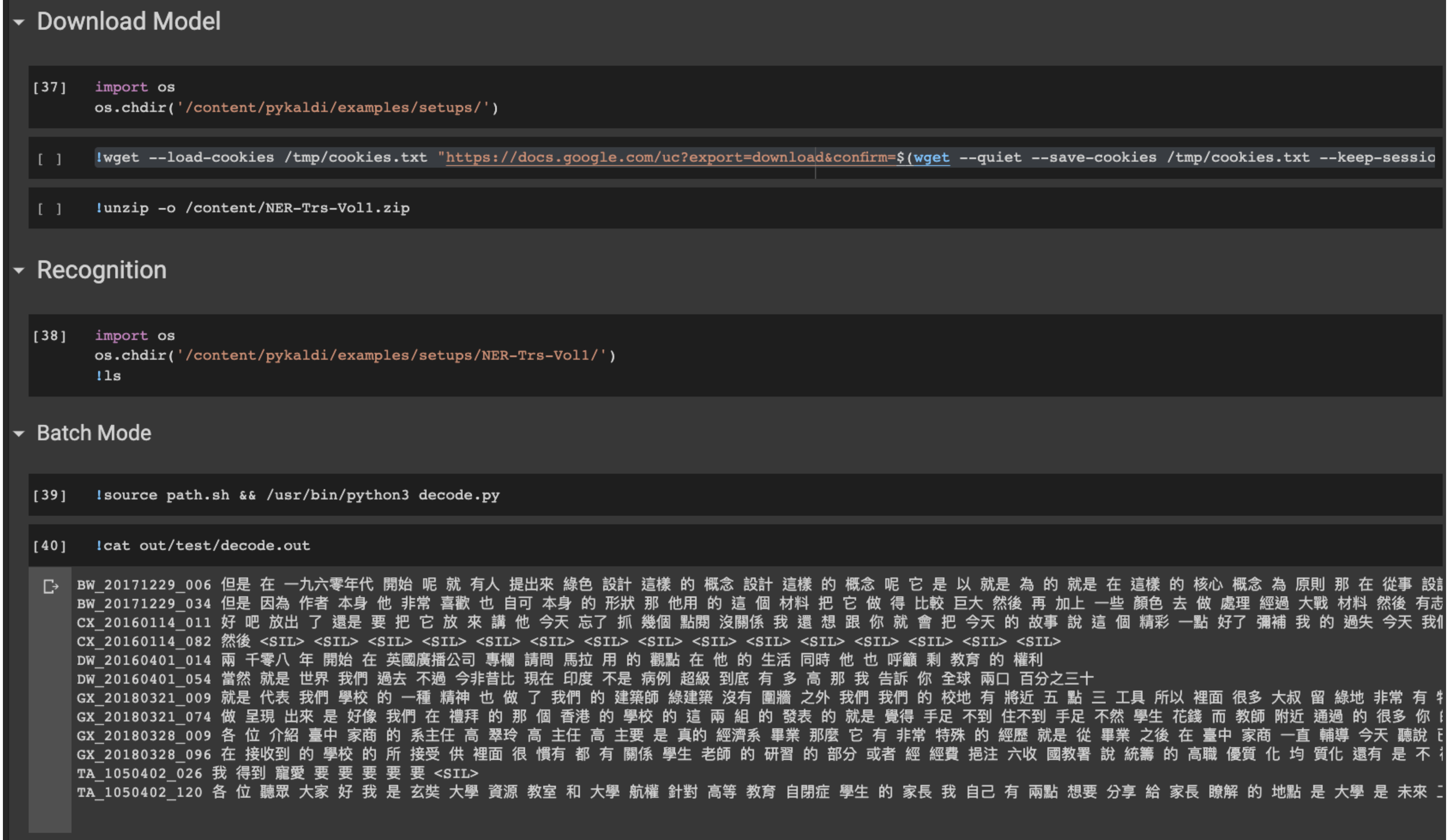
- model
  - final.dubm
  - final.ie
  - final.mat
  - final.mdl
  - global_cmvn.stats
  - HCLG.fst
  - words.txt



```
# Construct recognizer
decoder_opts = LatticeFasterDecoderOptions()
decoder_opts.beam = 30
decoder_opts.max_active = 7000
decoder_opts.min_active = 1000
#decoder_opts.ivector_scale = 1.0
decoder_opts.lattice_beam = 8
decodable_opts = NnetSimpleComputationOptions()
decodable_opts.acoustic_scale = 1.0
decodable_opts.frame_subsampling_factor = 3
decodable_opts.frames_per_chunk = 50
asr = NnetLatticeFasterRecognizer.from_files("model/final.mdl", "model/HCLG.fst", "model/words.txt", decoder_opts

# Define feature pipelines as Kaldi rspecifiers
feats_rspec = (
    "ark:compute-mfcc-feats --config=conf/mfcc_hires.conf scp:data/test/wav.scp ark:- |"
)
pitch_rspec = (
    "ark:compute-kaldi-pitch-feats --config=conf/pitch.conf --scp:data/test/wav.scp ark:- | process-kaldi-pitch-f
)
combi_rspec = (
    "ark:paste-feats 'ark:compute-mfcc-feats --config=conf/mfcc_hires.conf scp:data/test/wav.scp ark:- |' 'ark:co
)
ivectors_rspec = (
    "ark:compute-mfcc-feats --config=conf/mfcc_hires.conf scp:data/test/wav.scp ark:- |"
    "ivector-extract-online2 --config=conf/ivector_extractor.conf ark:data/test/spk2utt ark:- ark:- |"
)

# Decode wav files
with SequentialMatrixReader(combi_rspec) as f, \
     SequentialMatrixReader(ivectors_rspec) as i, \
     open("out/test/decode.out", "w") as o:
    for (key, feats), (_, ivectors) in zip(f, i):
        out = asr.decode((feats, ivectors))
        print(key, out["text"], file=o)
```



Download Model

```
[37]  import os
      os.chdir('/content/pykaldi/examples/setups/')
```

```
[ ]   !wget --load-cookies /tmp/cookies.txt "https://docs.google.com/uc?export=download&confirm=$(wget --quiet --save-cookies /tmp/cookies.txt --keep-sessio
```

```
[ ]   !unzip -o /content/NER-Trs-Vol1.zip
```

Recognition

```
[38]  import os
      os.chdir('/content/pykaldi/examples/setups/NER-Trs-Vol1/')
      !ls
```

Batch Mode

```
[39]  !source path.sh && /usr/bin/python3 decode.py
```

```
[40]  !cat out/test/decode.out
```

```
BW_20171229_006 但是 在 一九六零年代 開始 呢 就 有人 提出來 綠色 設計 這樣 的 概念 設計 這樣 的 概念 呢 它 是 以 就是 為 的 就是 在 這樣 的 核心 概念 為 原則 那 在 從事 設
BW_20171229_034 但是 因為 作者 本身 他 非常 喜歡 也 自可 本身 的 形狀 那 他用 的 這 個 材料 把 它 做 得 比較 巨大 然後 再 加上 一些 顏色 去 做 處理 經過 大戰 材料 然後 有志
CX_20160114_011 好 吧 放出 了 還是 要 把 它 放 來 講 他 今天 忘了 抓 幾個 點閱 沒關係 我 還 想 跟 你 就 會 把 今天 的 故事 說 這 個 精彩 一點 好了 彌補 我 的 過失 今天 我
CX_20160114_082 然後 <SIL> <SIL> <SIL> <SIL> <SIL> <SIL> <SIL> <SIL> <SIL> <SIL> <SIL> <SIL> <SIL> <SIL> <SIL>
DW_20160401_014 兩 千零八 年 開始 在 英國廣播公司 專欄 請問 馬拉 用 的 觀點 在 他 的 生活 同時 他 也 呼籲 剩 教育 的 權利
DW_20160401_054 當然 就是 世界 我們 過去 不過 今非昔比 現在 印度 不是 病例 超級 到底 有 多 高 那 我 告訴 你 全球 兩口 百分之三十
GX_20180321_009 就是 代表 我們 學校 的 一種 精神 也 做 了 我們 的 建築師 綠建築 沒有 圍牆 之外 我們 我們 的 校地 有 將近 五 點 三 工具 所以 裡面 很多 大叔 留 綠地 非常 有
GX_20180321_074 做 呈現 出來 是 好像 我們 在 禮拜 的 那 個 香港 的 學校 的 這 兩 組 的 發表 的 就是 覺得 手足 不到 住不到 手足 不然 學生 花錢 而 教師 附近 通過 的 很多 你
GX_20180328_009 各 位 介紹 臺中 家商 的 系主任 高 翠玲 高 主任 高 主要 是 真的 經濟系 畢業 那麼 它 有 非常 特殊 的 經歷 就是 從 畢業 之後 在 臺中 家商 一直 輔導 今天 聽說
GX_20180328_096 在 接收到 的 學校 的 所 接受 供 裡面 很 慣有 都 有 關係 學生 老師 的 研習 的 部分 或者 經 經費 挹注 六收 國教署 說 統籌 的 高職 優質 化 均 質化 還有 是 不
TA_1050402_026 我 得到 寵愛 要 要 要 要 要 <SIL>
TA_1050402_120 各 位 聽眾 大家 好 我 是 玄奘 大學 資源 教室 和 大學 航權 針對 高等 教育 自閉症 學生 的 家長 我 自己 有 兩點 想要 分享 給 家長 瞭解 的 地點 是 大學 是 未來
```

CF: Standalone Mode (1/2)

```
steps/online/nnet3/prepare_online_decoding.sh --add_pitch true
data/lang_chain/ exp/nnet3/extractor exp/chain/tdnn_1a_sp
exp/chain/nnet_online
```

CF: Standalone Mode (2/2)

```
#!/usr/bin/env bash
. ./path.sh

online2-wav-nnet3-latgen-faster --config=conf/online.conf --add-pitch=true \
--do-endpointing=false --frames-per-chunk=50 --extra-left-context-initial=0 \
--online=true --frame-subsampling-factor=3 --max-active=7000 \
--min-active=1000 --beam=15.0 --lattice-beam=8.0 \
--online=false --acoustic-scale=1.0 \
--word-symbol-table=model/words.txt model/final.mdl model/HCLG.fst \
ark:data/test/spk2utt scp:data/test/wav.scp \
ark,t:out/test/standalone-decode.txt
```