pyBK

audio, speech activity detection (SAD) and unpartitioned evaluation map (UEM)

UEM: Along with all audio ﬁles an index ﬁle will be distributed indicating the segments within the audio ﬁlesthat are part of the evaluation. The format of these ﬁles will be NIST Unpartitioned Evaluation Map

UEM: source ID(filename), channel, evaluation segment start time, evaluation segment end time

Each line in the UEM indicates a segment in an audio file that must be processed by the speech recognition system

segmentTable=getSegmentTable(mask,speechMapping,config.getint('SEGMENT','length'),config.getint('SEGMENT','increment'),config.getint('SEGMENT','rate'))

kbm, gmPool = trainKBM(data,config.getint('KBM','windowLength'),windowRate,kbmSize )

// output the result

getSegmentationFile

Acoustic features

Different speech sounds are generated by varying the shape of the vocal tract and its mode of excitation

Popular spectral features:

Mel Frequency Cepstral Coefficients (MFCC)

Linear Prediction Cepstral Coefficients (LPCC)

Perceptual Linear Prediction Cepstral (PLPC) Coefficients

Speech Activity Detection (SAD)

Energy-based: simple and fast, ineffective for noise like coughing, laughing etc..

Model-based: models for speech/non-speech, speech/music or noise etc;

Need to be trained with pre-labeled data

Speaker embedding

speaker-specific information

i-vector architectures

one to extract the i-vectors, and a second one to learn a probabilistic linear discriminant analysis (PLDA) scoring function todecide whether two i-vectors are from the same speaker or not.

Speech segmentation: fixed length sliding window 🡪 smaller nonoverlapping segments 🡪 fixed length embeddings 🡪 clustering

Naive online clustering: cosine similarity, threshold

Audio embedding extraction: speaker-discriminative embeddings: specific features such as MFCCs , speaker factors, i-vectors, d-vectors

i-vectors, d-vectors: speaker recognition, verification

d-vectors: deep neural network based embedding; neural networks can be trained with big datasets

Clustering: unsupervised previously

Gaussian mixture models,

Agglomerative hierarchical clustering

k-means

Links

Spectral clustering

Uis-rnn: replace the unsupervised clustering module by

an online generative process that naturally incorporates labelled data

for training.

(1) Each speaker is modeled by an instance of RNN, and these instances share the same parameters;

(2) An unbounded number of RNN instances can be generated

(3) The states of different RNN instances, corresponding to different speakers

Moreover, instead of using manually hand-crafted spectral features such as MFCC [9], PLP [10] or RASTA [11], we apply a recurrent neural network directly on the magnitude

spectrograms (SFTF) to learn a set of high-level feature representations also referred to as speaker embeddings. Our motivation to use SFTF and not raw audio stream stems from the fact that deep architectures were successfully used to learn Mel-like filters from power spectrum [12], while applying them directly on raw audio data did not lead to classification performance improvement

An i-vector is an information-rich low-dimensional vector extracted from a feature sequence that represents a speech segment, sometimes known as an audio voiceprint per analogy to fingerprints

An i-vector is an information-rich low dimensional vector extracted from a feature sequence that represents a speech segment, sometimes known as an audio voiceprint per analogy to fingerprints

They create a representation of speech utterance from a set of pre-trained speaker models in order to get a likelihood score of each anchor. Those representations are then grouped into characterization vectors that describe individual speakers. The anchor modeling allows to embed the information about the speaker into the speech segment representation and is often referred to as a speaker embedding model. Similar embeddings have recently been obtained through deep neural network training [24]. More precisely, a neural network with three hidden layers was applied to Gaussian Mixture Universal Background Model (GMM-UBM) of MFCC features with the goal of speaker classification. The activations of the so-trained neural network were then used as features of previously unseen speakers

State-of-the-art systems represent segments by making use of a Gaussian mixture model (GMM) adapted from the UBM to form an i-vector