# Speaker diarization

Steve Renals

Automatic Speech Recognition – ASR Lecture 18
21 March 2019

# Speaker recognition

- Speaker identification – determine which of the set of enrolled speakers a test speaker matches
- Speaker verification – determine if a test speaker matches a *specific speaker*
- Speaker diarization – "who spoke when" segment and label a continuous recording by speaker

# Speaker recognition

- Speaker identification – determine which of the set of enrolled speakers a test speaker matches
- **Speaker verification** (last lecture) – determine if a test speaker matches a *specific speaker*
- **Speaker diarization** – "who spoke when" segment and label a continuous recording by speaker

# Speaker diarization

# Dealing with multiple speakers

- Speaker diarization is the "who spoken when" task: given a recording, divide it into segments, where each segment corresponds to speech of a single speaker
- Each recording contains multiple speakers – unlike what we have assumed so far for speech recognition and speaker verification
- Multiple speakers in a recording is realistic – many possible domains, e.g.:
    - Broadcast media
    - Telephone conversations
    - Call centres
    - Meeting recordings

# Dealing with multiple speakers

- Speaker diarization is the “who spoken when” task: given a recording, divide it into segments, where each segment corresponds to speech of a single speaker
- Each recording contains multiple speakers – unlike what we have assumed so far for speech recognition and speaker verification
- Multiple speakers in a recording is realistic – many possible domains, e.g.:
  - Broadcast media
  - Telephone conversations
  - Call centres
  - Meeting recordings
- A basic approach to diarization:
  Segment the recording into a sequence of short pieces, each assumed to be a single speaker. Then treat as a speaker verification task between all pairs of segmented utterances
  - Guaranteed to fail on segments with overlapping speakers!

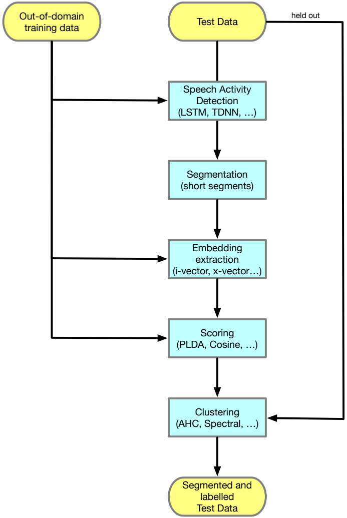
# Measuring speaker diarization – Diarization error rate

- There are three main type of error to consider in speaker diarization:
  - **Missed speech** ($E_{\text{miss}}$): system labels a segment as non-speech, but segment is attributed to a speaker in the reference
  - **False-alarm speech** ($E_{\text{fa}}$): system attributes segment to a speaker, but segment is labelled as non-speech in the reference
  - **Speaker error** ($E_{\text{spkr}}$): system attributes segment to a speaker different to the reference attribution
- These errors are computed in a time-based way: each is expressed as a fraction of the scored time in the reference
- The diarization error rate ($DER$) is computed as a sum of these errors

$$DER = E_{\text{miss}} + E_{\text{fa}} + E_{\text{spkr}}$$

- Note that $E_{\text{miss}}$ and $E_{\text{fa}}$ arise from the speech activity detection

# Framework for speaker diarization


Out-of-domain training data
Test Data
held out
Speech Activity Detection (LSTM, TDNN, ...)
Segmentation (short segments)
Embedding extraction (i-vector, x-vector...)
Scoring (PLDA, Cosine, ...)
Clustering (AHC, Spectral, ...)
Segmented and labelled Test Data

Segment a recording, and attach a speaker label to each segment.

1. Split the recording into segments
2. Speech activity detection: identify whether each segment is speech or non-speech, discard non-speech
3. Represent the speech segments using some form of fixed length embedding: i-vector, x-vector, d-vector...
4. Compare all pairs of segments using a scoring metric such as PLDA
5. Cluster the segments using an algorithm such as agglomerative hierarchical clustering

# Segmentation and Speech Activity Detection

- Speech activity detection (SAD) typically carried out using an LSTM or TDNN neural network trained on a large amount of diverse data
  - Binary output: speech vs. non-speech
  - Possibly with data augmentation – noise, reverb, etc.
- Following SAD, segment into short fixed-length segments (typically 2s)
  - Assumes each segment contains speech from a single speaker
  - In practice can use overlapping segments (overlap by 0.5s at start and end)
  - Relatively short segment duration for embedding computation

# Speaker Embeddings and Clustering

- Compute a speaker representation for each segment
    - i-vector - typically 64-128 dimension
    - x-vector / d-vector - typically 128-256 dimension
    - can reduce the dimension by performing PCA on the set of embeddings for a recording
- Score all segment pairs – typically use PLDA
- Cluster segments – many possible clustering algorithms: Agglomerative hierarchical clustering can work well
    - Only need to compute pairwise segment scores once
    - Score for a cluster pair is obtained by averaging the pairwise scores between the segments in each cluster
- Determine the number of clusters
    - Clustering stopping criterion determines the number of clusters
    - Define a prior distribution on the number of speakers, and apply to clustering
    - Bayesian models with a prior on number of clusters – Variational Bayes (VB) HMM, Hierarchical Dirichlet Process (HDP) HMM, distance-dependent Chinese Restaurant Process (ddCRP), . . .

# DIHARD

- R&D in speaker diarization has been very domain-dependent
  - 1990s – broadcast news (Hub4)
  - 2000s – multi-microphone meeting recordings (AMI, NIST RT)
  - 2010s – conversational telephone speech (CallHome)
- Had the effect of fragmenting the field
- Since 2018 the DIHARD Challenge (`https://coml.lscp.ens.fr/dihard/`) has focused on "speaker diarization for challenging recordings where there is an expectation that the current state-of-the-art will fare poorly" – diverse set of data sets used

# Some hot topics in diarization

- Overlapping speech – most systems do not explicitly deal with this
- Speech activity detection is still a significant cause of error
- Development of end-to-end systems
- Bayesian approaches (learning the number of speakers/clusters from the data)
- Use of supervised learning

# Reading

- D Garcia-Romero et al (2017), "Speaker diarization using deep neural network embeddings", ICASSP. https://ieeexplore.ieee.org/document/7953094
- G Sell et al (2018), "Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge", Interspeech. https://www.isca-speech.org/archive/Interspeech_2018/abstracts/1893.html
- K Church et al (2017), "Speaker diarization: A perspective on challenges and opportunities from theory to practice", ICASSP. https://ieeexplore.ieee.org/abstract/document/7953098