Speech Enhancement and Diarization

JSALT 2024 Summer School

Matthew Maciejewski

June 12, 2024

Johns Hopkins University

or, towards:

The Cocktail Party Problem[1]

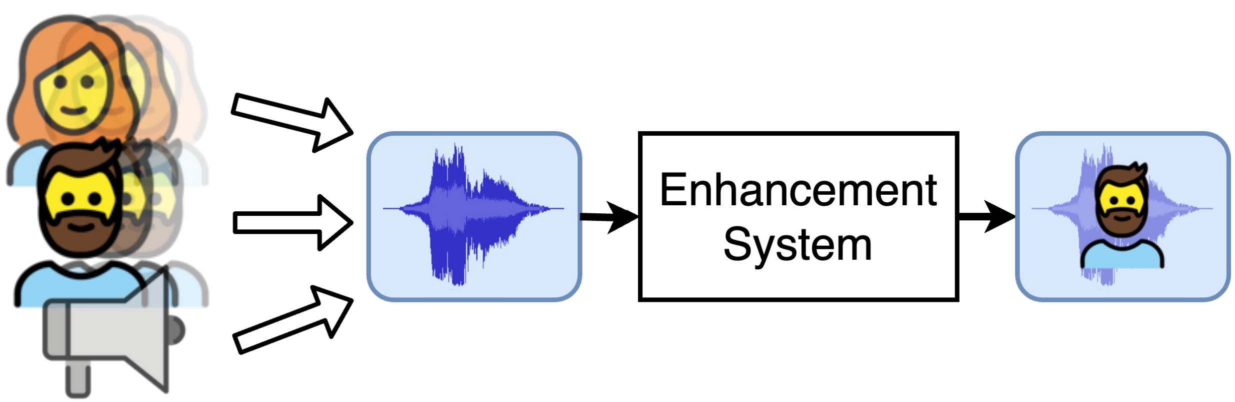
[1]E.C. Cherry, *Some Experiments on the Recognition of Speech, with One and with Two Ears.* The Journal of the Acoustical Society, 1953

Speech Enhancement (and Separation)

What is speech enhancement?

- Recordings of speech often have a lot of degradation and interfering sounds

- Speech enhancement is the task of removing interferences or reconstructing the clean speech



Why do we care?

- Human listening can always be the end goal
- Degraded audio often leads to degraded performance of downstream systems
- Robust speech technology often integrates techniques developed in enhancement

Mathematical Formulation

Input: $x(t) = s(t) + n(t)$

Output: $y(t) = \hat{s}(t)$

We can also treat $n(t)$ more precisely:

Reverberation: $x(t) = s(t) * h_{RIR}(t)$

Separation: $x(t) = s_1(t) + s_2(t)$

All together:

$$x(t) = \sum_{c=1}^{C} [s_c(t) * h_c(t)] + \sum_{k=1}^{K} n_k(t)$$

Performance Evaluation

- Full Reference
  - **SI-SDR**, SNR, (SDR, SIR, SAR), …
  - **PESQ**, STOI, POLQA, …
- No Reference
  - Human listening tests! (**MOS**)
  - ITU P.563, SRMRnorm, …
  - DNSMOS, SQAPP, …
- Downstream Evaluation
  - Impact on downstream speech tasks

$$\mathrm{SI{-}SDR} = 10\log_{10}\frac{|s|^2}{|s-\beta\hat{s}|^2}$$

$$\text{for } \beta \text{ s.t. } s \perp s-\beta\hat{s}$$

Significance of Ground Truth

Issues of ground truth are a significant aspect of waveform-level tasks

- Non-full-reference metrics have large downsides, full-reference (typically) require synthetic mixtures
- Neural network training targets typically require targets and also require synthetic mixtures
  - Domain mismatch can be a significant problem
- Practical approaches often avoid trying to directly optimize the output waveform

JOHNS HOPKINS UNIVERSITY

General Approach

- Speech enhancement methods generally fall under the umbrella of "filtering", with some further broad categorizations:

temporal filtering vs. **spectral filtering**

estimation vs. **decomposition**

- These distinctions are in some sense arbitrary and can often be considered equivalent

Mask-Based Enhancement

classification →

regression →

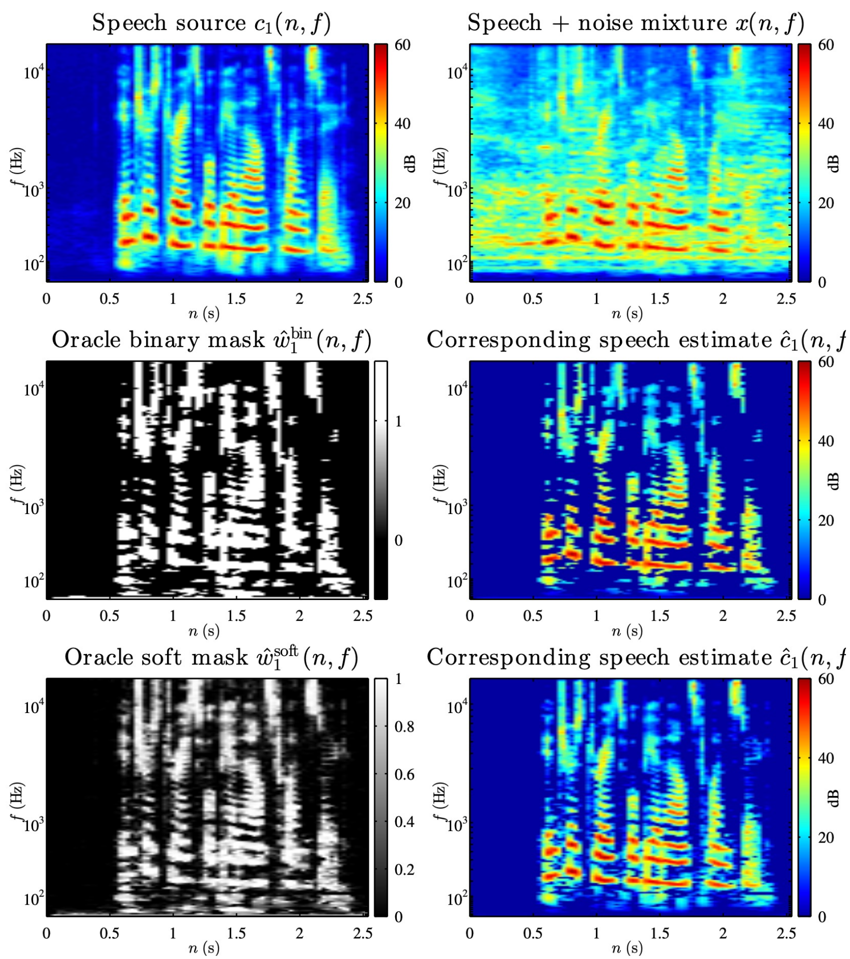


 Image credit: Vincent, et al. *Audio Source Separation and Speech Enhancement*

How do we estimate the filters?

- Can be learning-free, unsupervised, supervised
- Estimation of speech presence probability, noise distribution, SNR, power spectra, etc.
- Nonnegative Matrix Factorization (NMF)
  - Decompose magnitude/power spectrum into set of distinct basis spectra
- Independent Component Analysis (ICA)
  - Assumes mixture of mutually-independent stochastic source signals

Reverberation

- Room Impulse Response (RIR) captures room reflections and mixes via convolution

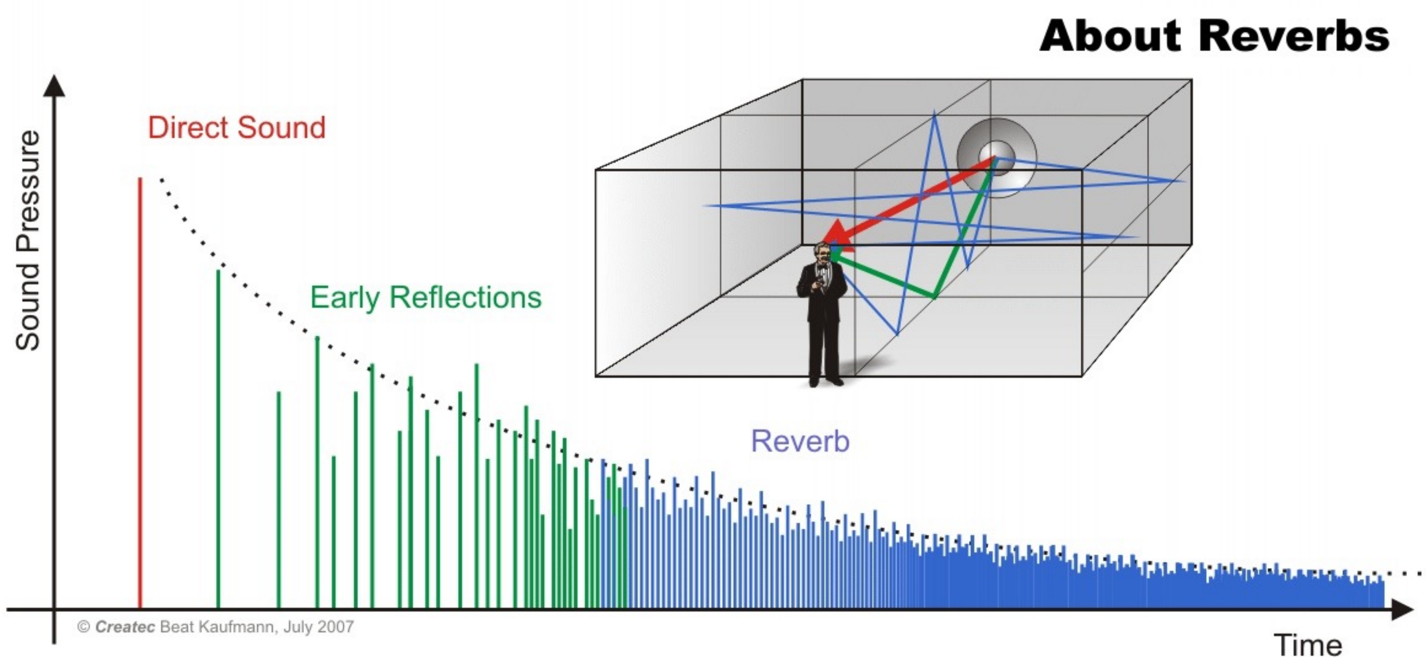


Image credit: *Createc* Beat Kaufmann

Spectral Effect of Reverberation

- Reverb results in spectral smearing

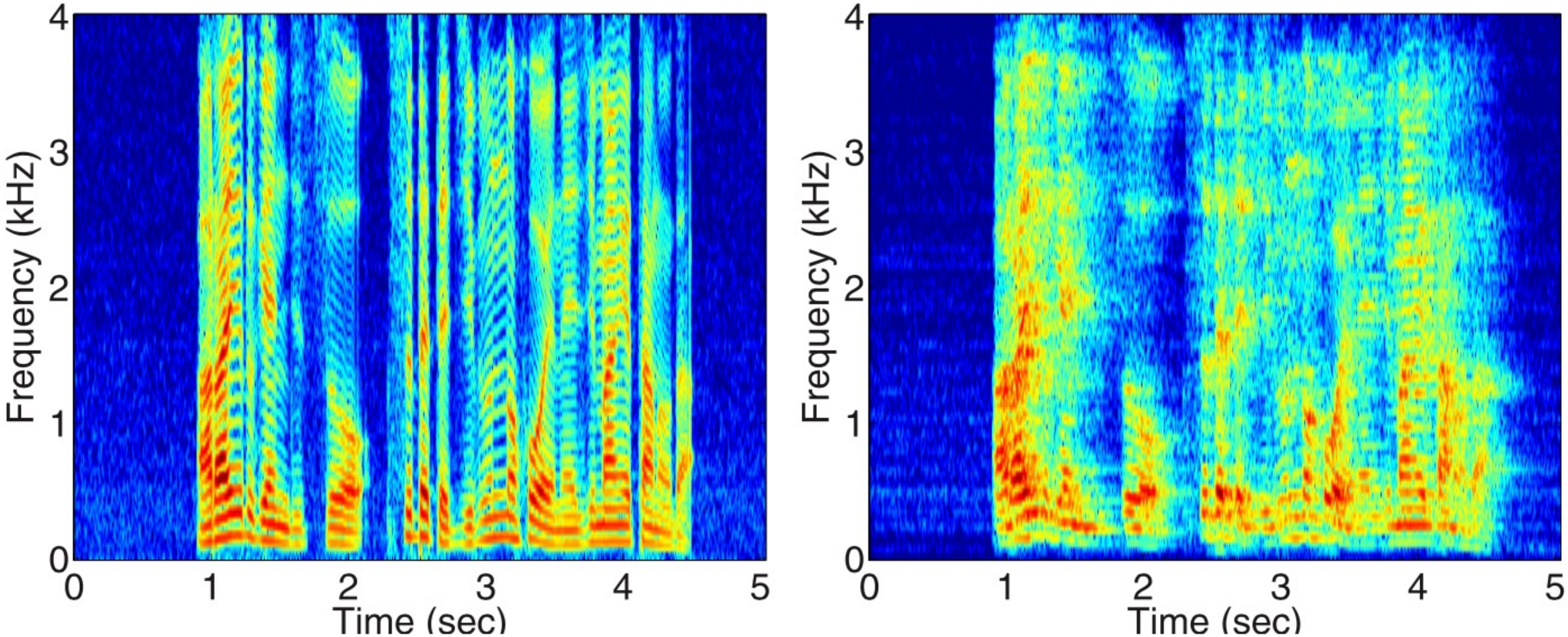


Image credit: Cohen, et al. *Speech Processing in Modern Communication*

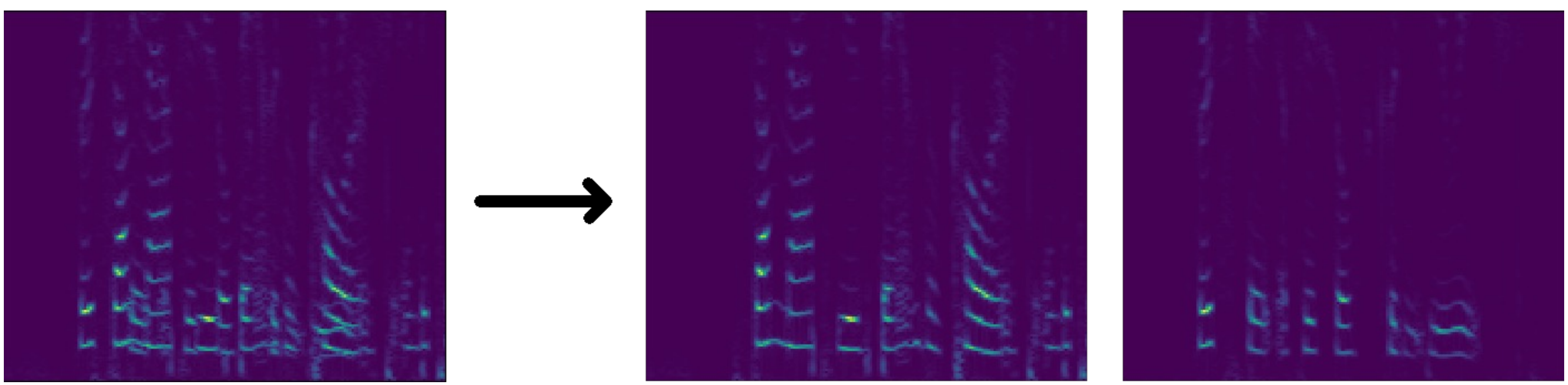
De-Reverberation

- Most successful practical approach is Weighted Prediction Error (WPE)[1,2] dereverberation
- The late tail reverberation is estimated and cancelled via delayed linear prediction
  - Iterative procedure to continually update inverse filter
- Avoiding early reflections minimizes corruption of direct path and issue of relative non-stationarity
- "Deep" extension via neural speech Power Spectral Density (PSD) estimation[3]

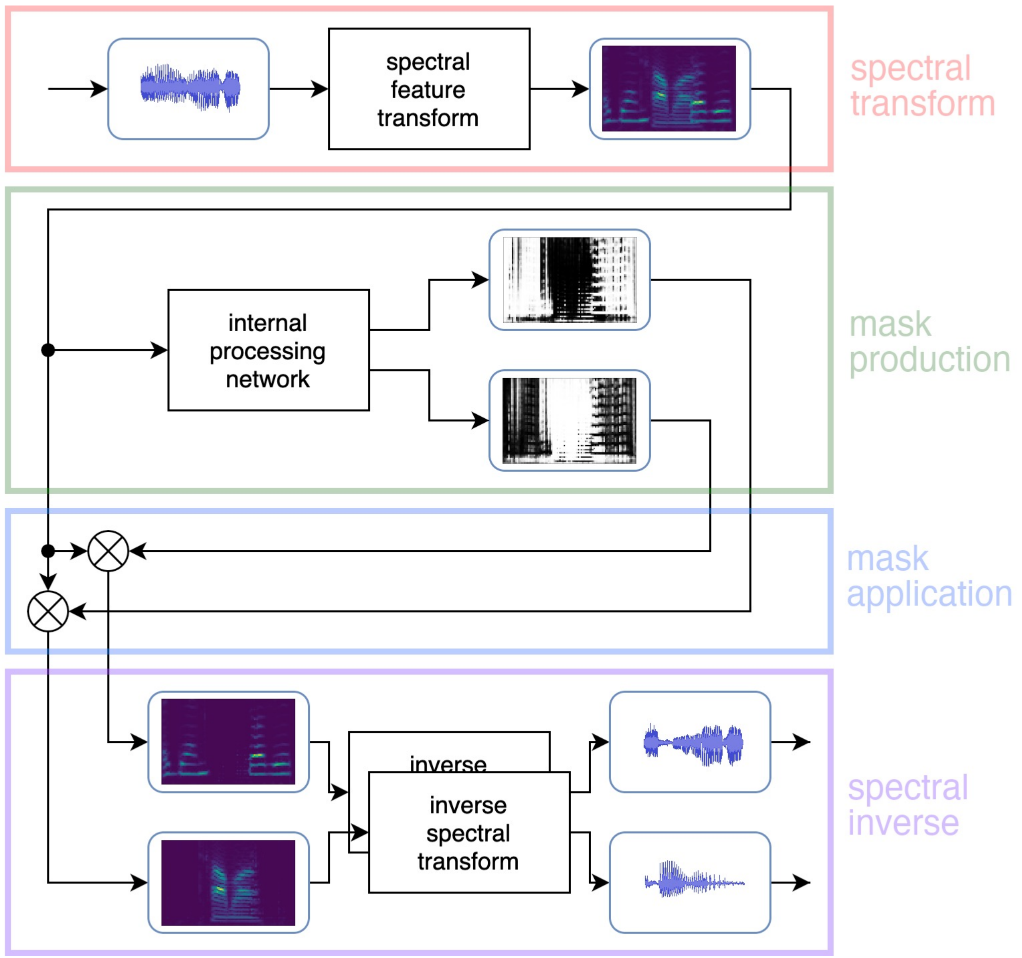
[1]T. Nakatani (2010) [2]T. Yoshioka (2012) [3]K. Kinoshita (2017)

Speech Separation

- Speech separation aims to estimate single-speaker waveforms from overlapping speech
- Relies on the spectral sparsity of speech



Separation Pipeline



JOHNS HOPKINS
UNIVERSITY

Challenges in Training

Foundational approaches on mask-based loss:

- Deep Clustering (DPCL)
  - Extract embedding for each STFT bin
  - Ensure self-similarity of dominant bins from a speaker
- Permutation-Invariant Training (PIT)
  - Compute minimum loss across all output permutations, backpropagate from best permutation

- State-of-Art systems dominated by learned spectral transforms with SI-SDR PIT loss

[1]J.R. Hershey (2016) [2]D. Yu (2016)

Johns Hopkins University

Target Speaker Extraction

- Given a recording and an enrollment utterance or speaker representation, produce the clean speech of the enrolled speaker

- Has elements of both speech separation and speech enhancement

Multichannel Enhancement

- Collecting audio simultaneously with multiple microphones gives more information for the underlying signals

- Particularly: multiple sensors allows for localization, and multiple sources generally have different locations

Formulation

$$\boldsymbol{x}(t) = \sum_{j=1}^{J} \boldsymbol{c}_j(t)$$

diffuse source → just $\boldsymbol{c}_j(t)$

point source → (time−invariant) spatialization

$$\boldsymbol{c}_j(t) = \boldsymbol{a}_j(t) * s_j(t)$$

$\boldsymbol{x} \in \mathbb{R}^{I\times T}$
I microphones

$\boldsymbol{c} \in \mathbb{R}^{I\times T}$
spatialized sources

$$\boldsymbol{a}_j(t) = \left[a_{1j}(t), \dots, a_{Ij}(t)\right]^{\mathrm{T}}$$

can be RIR, delay/attenuation

Can approximate in STFT domain:

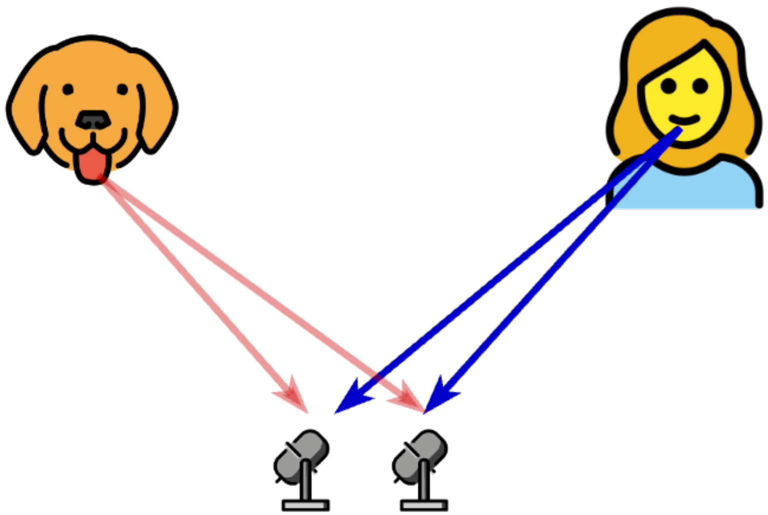
$$\boldsymbol{a}_j(n,f) \sim \boldsymbol{a}_j(f)$$
$$\boldsymbol{c}_j(n,f) = \boldsymbol{a}_j(f) s_j(n,f)$$
$$\boldsymbol{x}(n,f) = \boldsymbol{A}(f)\boldsymbol{s}(n,f)$$

"steering vector":
$\boldsymbol{a}_j(f) \rightarrow \boldsymbol{d}_j(f)$

$$\boldsymbol{d}_j(f) = \begin{bmatrix} \frac{1}{\sqrt{4\pi} r_{1j}} e^{-2j\pi r_{1j} v_f / c} \\ \vdots \\ \frac{1}{\sqrt{4\pi} r_{Ij}} e^{-2j\pi r_{Ij} v_f / c} \end{bmatrix} \approx \begin{bmatrix} e^{-2j\pi r_{1j} v_f / c} \\ \vdots \\ e^{-2j\pi r_{Ij} v_f / c} \end{bmatrix}$$

Beamforming

- “Delay and sum” beamforming aligns target signal temporally and misaligns other signals for constructive/destructive interference



TDOA Estimation

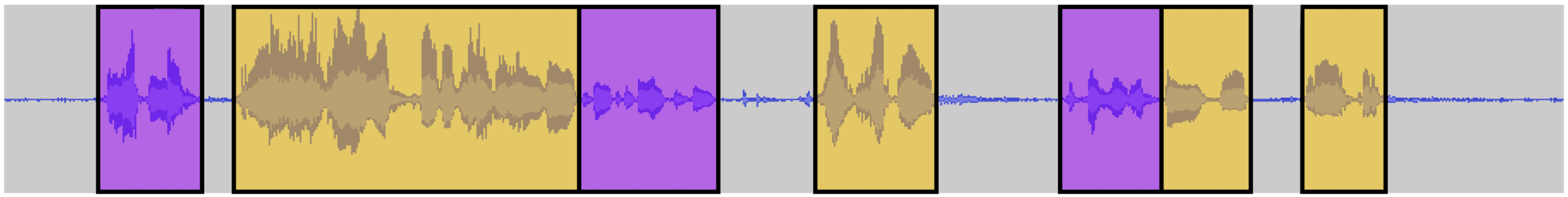
- Beamforming requires the "time difference of arrival" (TDOA)
- Generalized Cross-Correlation with Phase Transform (GCC-PHAT)[1]
- Minimum Variance Distortionless Response (MVDR) beamformer is computed in STFT domain by minimizing the power of the interfering signal
  - Weights can be computed from speech TF mask
  - Amenable to neural estimation

[1]C.H. Knapp (1976)

...questions?

(Speaker) Diarization

What is speaker diarization?



Who spoke when?

*other types of diarization exist, most notably language diarization

Why do we care?

- Many speech systems "malfunction" in multi-talker scenarios
    - Closed captioning or meeting transcription
    - Target speaker recognition

- Conversational analysis
    - Biomarkers for emotional state
    - Study of child language acquisition
    - Social role (e.g. interruptions)

Mathematical Formulation

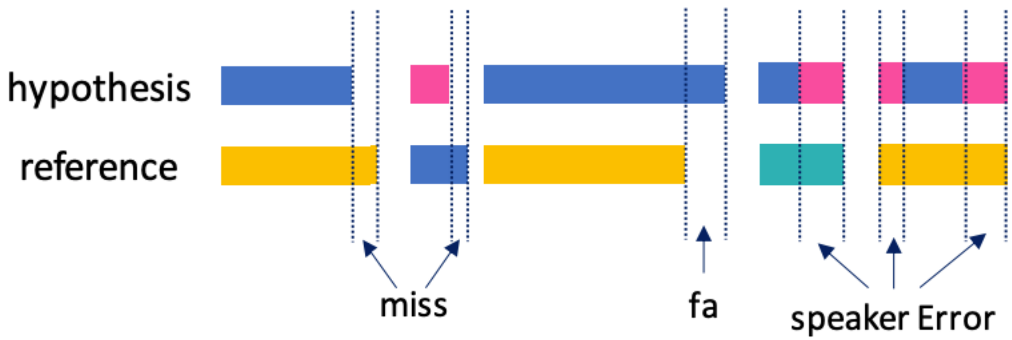
- “label-free” time series multi-label classification

$$x[t] \longrightarrow \boxed{\text{System}} \longrightarrow \boldsymbol{y}[t]$$

$$\boldsymbol{y} \in \{0, 1\}^{T \times S}$$

Order of speakers $s_i \in S$ does not matter

Metrics



- Diarization Error Rate (DER%)

$$\mathrm{DER} = \frac{false\_alarm + \ missed\_speech \ + \ speaker\_error}{total\_speech}$$

- Jaccard Error Rate (JER%)

$$\mathrm{JER} = \frac{1}{S}\sum_{i=1}^{S}\frac{false\_alarm_i + missed\_speech_i}{speech_i}$$

Image credit: Paola Garcia

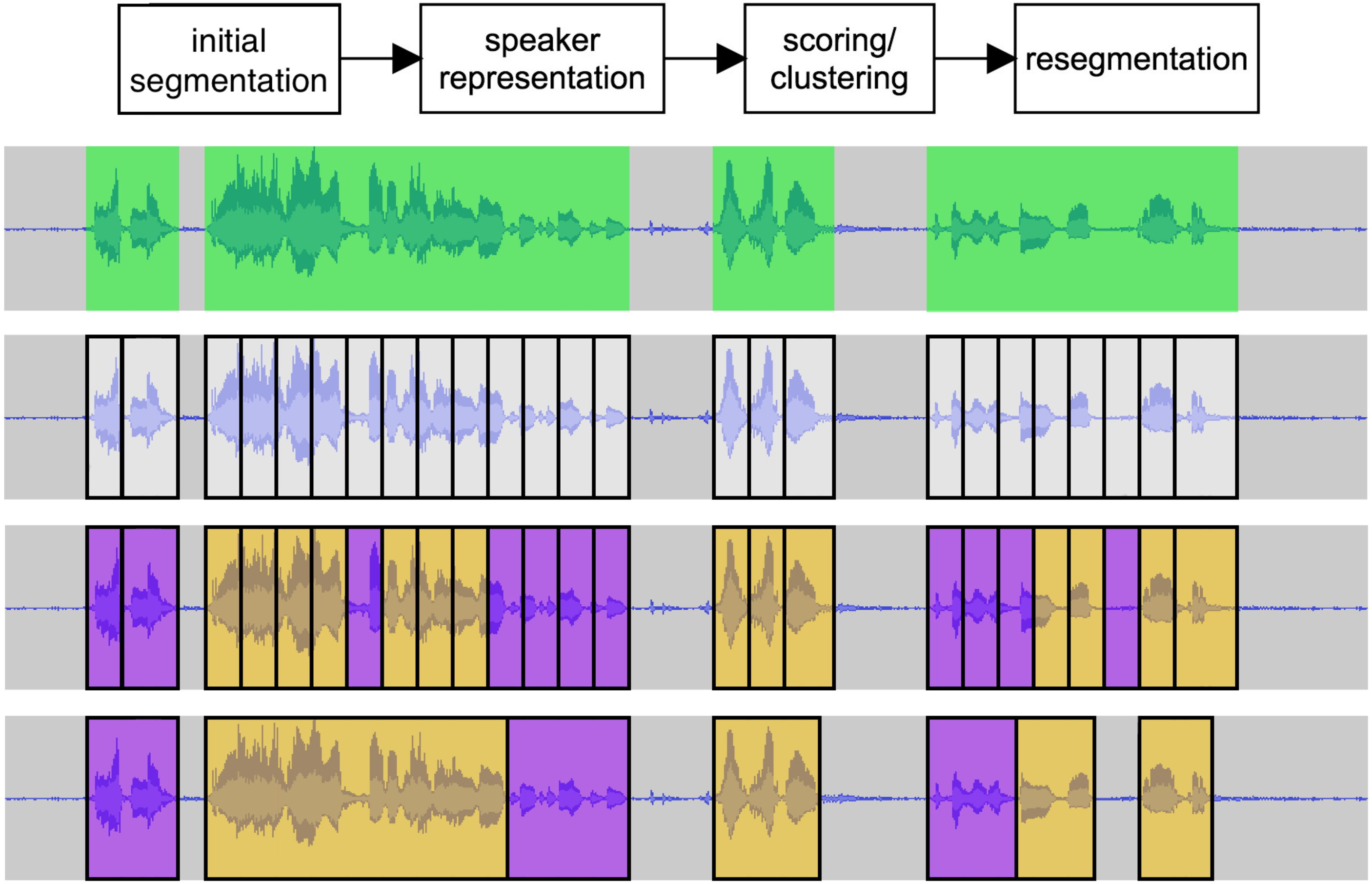
Johns Hopkins University

Approaches to Diarization

- Traditional “Clustering” Approaches
  - Multi-stage pipelines with independent components
  - Individually tuned
  - Less conducive to overlap detection

- Neural (End-to-End) Approaches
  - Trained to produce outputs directly
  - Can be jointly optimized
  - Resource intensive

Traditional Approach



JOHNS HOPKINS UNIVERSITY

Initial Segmentation

- Speech Activity Detection (SAD)
  - Basic speech presence classifier
  - Generally neural, statistical has been used

- Less commonly can be more sophisticated
  - Speaker change detection
  - Overlap detection

Speaker Representation

- Out-of-the-box Speaker ID systems
  - i-vectors, x-vectors, d-vectors
- Typically extracted under a sliding window
- Scoring can be tuned to test conditions or smaller speaker variability

Clustering

- Many clustering approaches
  - Agglomerative clustering
  - Spectral clustering

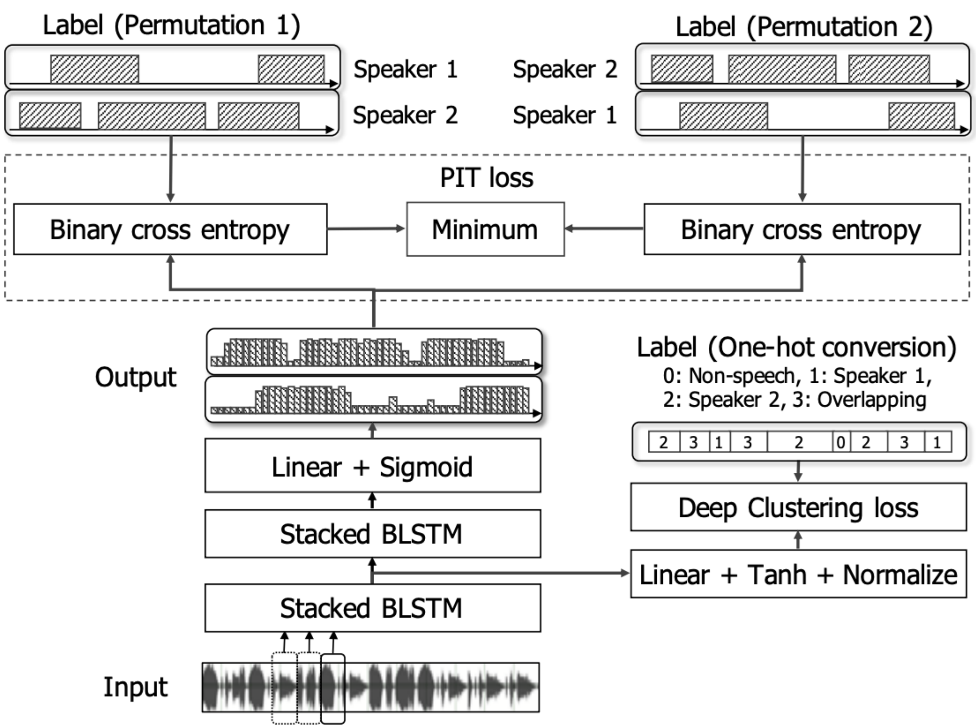
- Major challenge is speaker counting
  - Ground truth (not necessarily optimal!)
  - Speaker count estimation
  - Thresholding/Calibration

Resegmentation

- Variational Bayes HMM of x-vectors (VBx)
  - Probabilistic model treating x-vectors as observation of latent states corresponding to speakers
  - Models the temporal aspect of conversations
- Target Speaker Voice Activity Detection (TS-VAD)
  - Speaker-specific speech activity classifier based on input speaker representation
  - Handles overlap!

Neural Diarization

- Most methods derived from End-to-End Neural Diarization (EEND)[1] approach

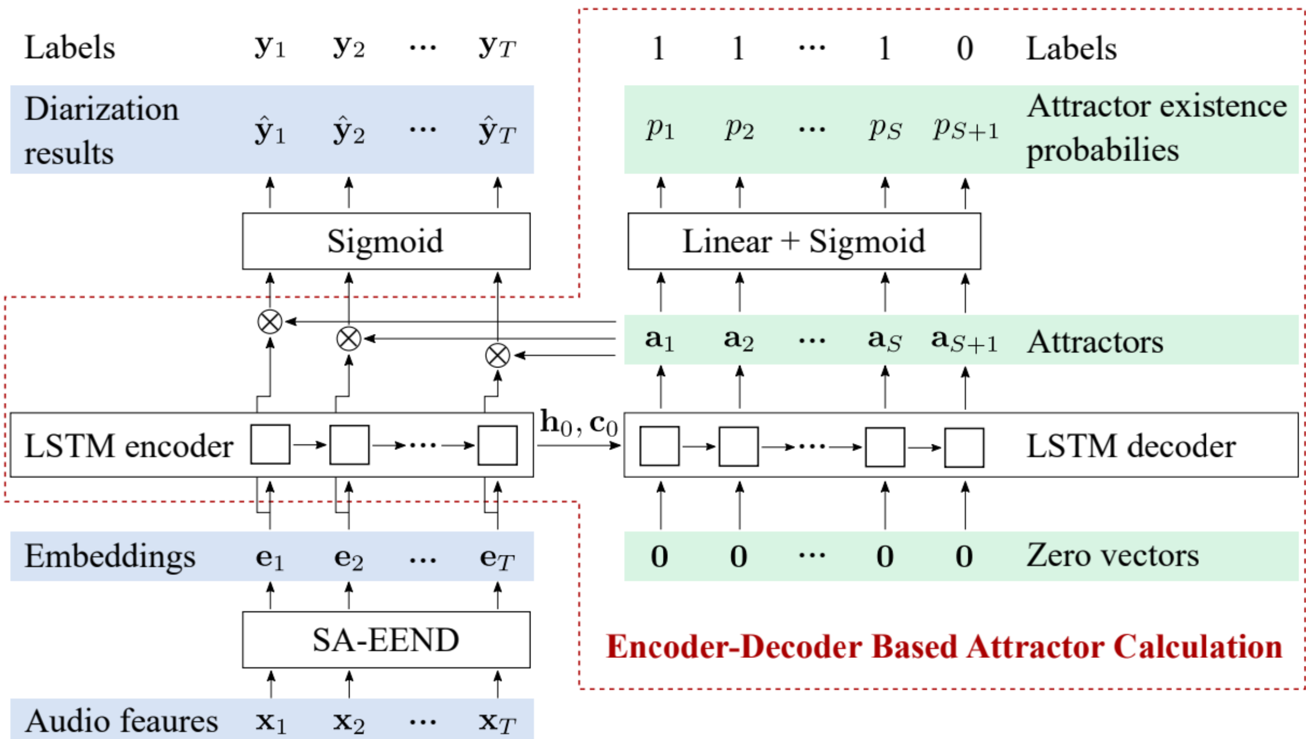


1

[1]Y. Fujita (2019)

Extension to Arbitrary Speakers

- Encoder-Decoder Attractors (EEND-EDA)[1] are used to model a variable number of speakers



[1]

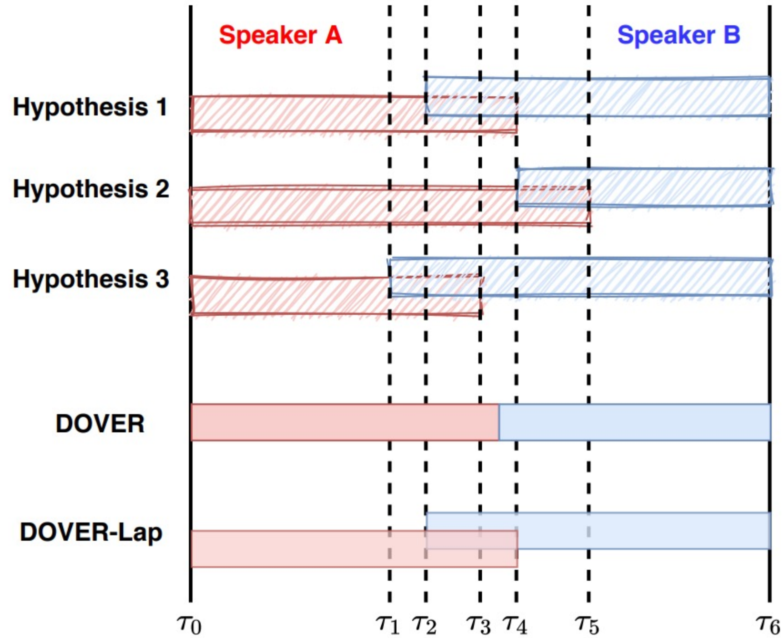
[1]S. Horiguchi (2020)

Practical Considerations

- Large amounts of data are required
- Memory requirements in training
  - Someone may talk long periods apart
- Processing long recordings
  - Must track speakers across block processing

System Ensembling

- Different systems may have different strengths and weaknesses (e.g. traditional vs. neural)
- DOVER[1] and overlap-aware extension DOVER-Lap[2]



2

[1]A. Stolcke (2019) [2]D. Raj (2021)

JOHNS HOPKINS UNIVERSITY

Multichannel Diarization

- Multiple microphones improve localization, and different talkers will be in different locations
  - They may, however, move around
- Directional information from beamforming may be integrated into the system
- Multiple audio signals may be used directly in the system, integrating beamforming implicitly

Multimodal Diarization

- Video may contain useful information for diarization and we would like to use it
- Audio-visual diarization has been successfully done using lip region of interest features[1]
  - Occlusions and out-of-frame issues pose a challenge

[1] He, et al. *End-to-End Audio-Visual Neural Speaker Diarization*

...questions?