Neural Models for Speaker Diarization: In the Context of Speech Recognition

Kyu J. Han, Director of Speech Modeling, ASAPP
Tae Jin Park, PhD Student, University of Southern California
Dimitrios Dimitriadis, Principal Researcher, Microsoft Research

Kyu Jeong Han

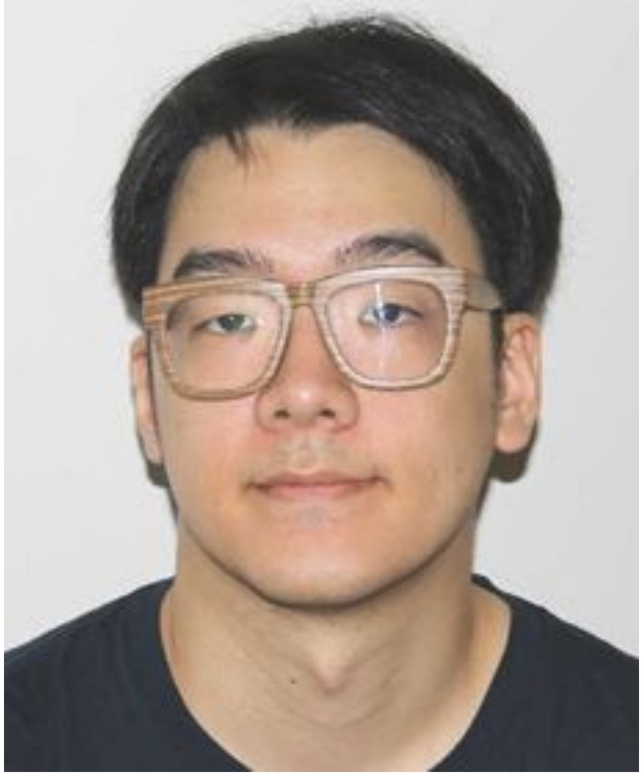
ASAPP

Kyu Jeong Han

Director of Speech Modeling at ASAPP

Received his PhD from USC in 2009 and is currently working for ASAPP Inc. leading deep learning technologies for speech applications in customer interaction domains. Dr. Han held research positions at IBM, Ford, Capio.ai (acquired by Twilio) and JD.com. He is actively involved in the speech community as well, serving as reviewers for IEEE, ISCA and ACL journals and conferences, and a Speech and Language Processing Technical Committee member for the IEEE SPS since 2019. He also serves for IEEE SLT-2020 as part of the Organizing Committee. In 2018, he won the ISCA Award for the Best Paper Published in Computer Speech & Language 2013-2017.

Tae Jin Park



Tae Jin Park

PhD Candidate at University of Southern California

Tae Jin Park received his B.S. degree in electrical engineering and M.S. degree in Electric Engineering and Computer Science from Seoul National University, Seoul, South Korea. in 2010 and 2012, respectively. In 2012, he joined Electrical and Telecommunication Research Institute (ETRI), Daejeon, South Korea, as a researcher. He is currently a Ph.D. candidate in Signal Analysis and Interpretation Laboratory (SAIL) at University of Southern California (USC). He is interested in machine learning and speech signal processing concentrating on speaker diarization.

Dimitrios Dimitriadis

Microsoft

Dimitrios Dimitriadis

Principal Researcher at Microsoft Research, WA

Dimitrios Dimitriadis worked as a Researcher in IBM Research, NY and AT&T Labs, NJ, and lecturer P.D 407/80 in School of ECE, NTUA, Greece. He is a Senior Member of IEEE. He was part of the Program Committee for the Multi-Learn’17 Workshop, and the Organizing Committee for IEEE SLT'18 and ICASSP'23. He has also served as session chair in multiple conferences. Dr. Dimitriadis has published more than 60 papers in peer-reviewed scientific journals and conferences with over 1500 citations. He received his PhD degree from NTUA in February 2005. His PhD Thesis title is "Non-Linear Speech Processing, Modulation Models and Applications to Speech Recognition". His major was in D.S.P. with Specialization in Speech Processing.

Interview Panel

- Andreas Stolcke (Amazon)
- Douglas Reynolds (MIT Lincoln Lab)
- Gakuto Kurata (IBM)
- Katrin Kirchhoff (Amazon)
- Miguel Jette (Rev.ai)
- Naoyuki Kanda (Microsoft)
- Paola Garcia (JHU)
- Quan Wang (Google)
- Shinji Watanabe (JHU)
- Shri Narayanan (USC)
- Sriram Ganapathy (IISC)
- Xavier Anguera (ELSA)
- Yifan Gong (Microsoft)

Outlines

Chapter 1: Diarization Overview

Part 1: Introduction
Part 2: Speaker Diarization Pipeline
Part 3: Future of Speaker Diarization

Chapter 2: Speaker Diarization and ASR

Part 1: Speaker diarization enhanced by ASR outputs
Part 2: Lexical information used in speaker diarization
Part 3: Joint modeling of speaker diarization and ASR

Chapter 3: Challenges and the State of Speaker Diarization

Part 1: Challenges in speaker diarization
Part 2: The State of speaker diarization

Chapter 1

Diarization Overview

Chapter 1

1. **Part 1: Introduction**

   1.1. Introduction to Speaker Diarization
   1.2. Applications of Speaker Diarization

2. **Part 2: Speaker Diarization Pipeline**

   2.1. Speaker Embedding Extraction
   2.2. Clustering and Speaker Counting
   2.3. Modular Systems VS End-to-end Systems
   2.4. Diarization Evaluation

3. **Part 3: Future of Speaker Diarization**

   3.1. Human Listener vs Speaker Diarization
   3.2. Next level Diarization Technology

Chapter 1

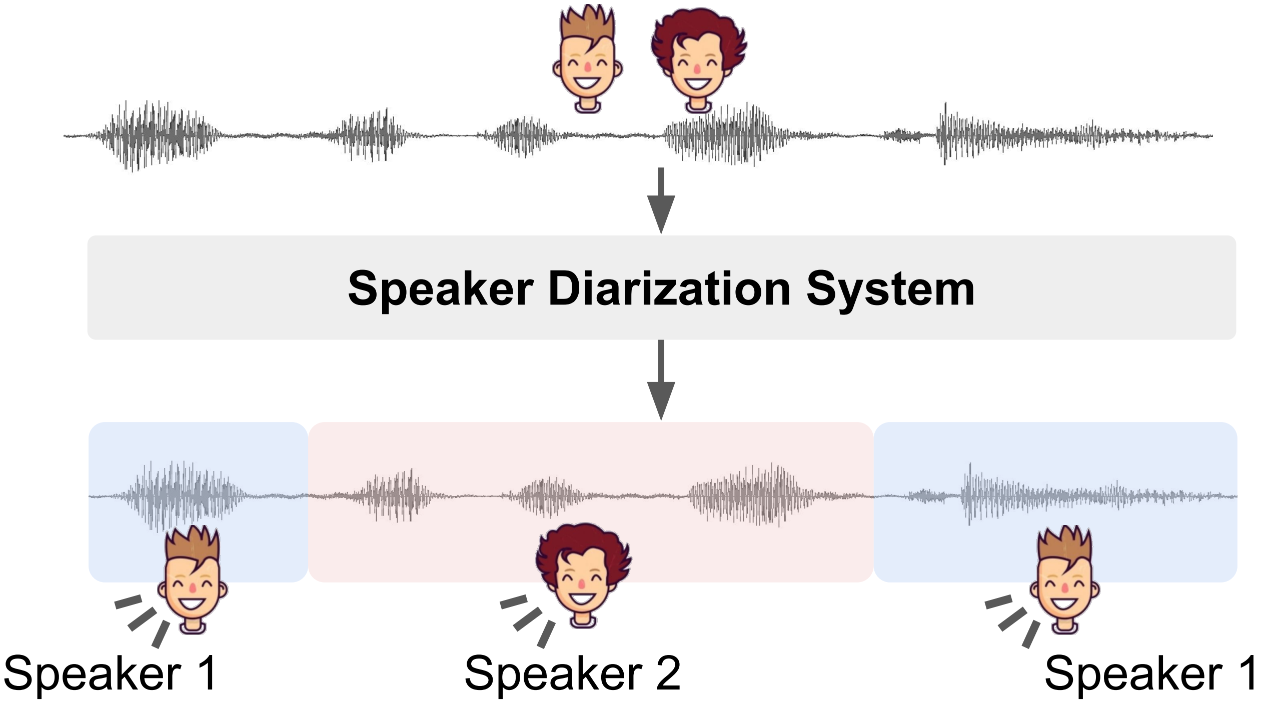
Diarization Overview

Part-1

Introduction

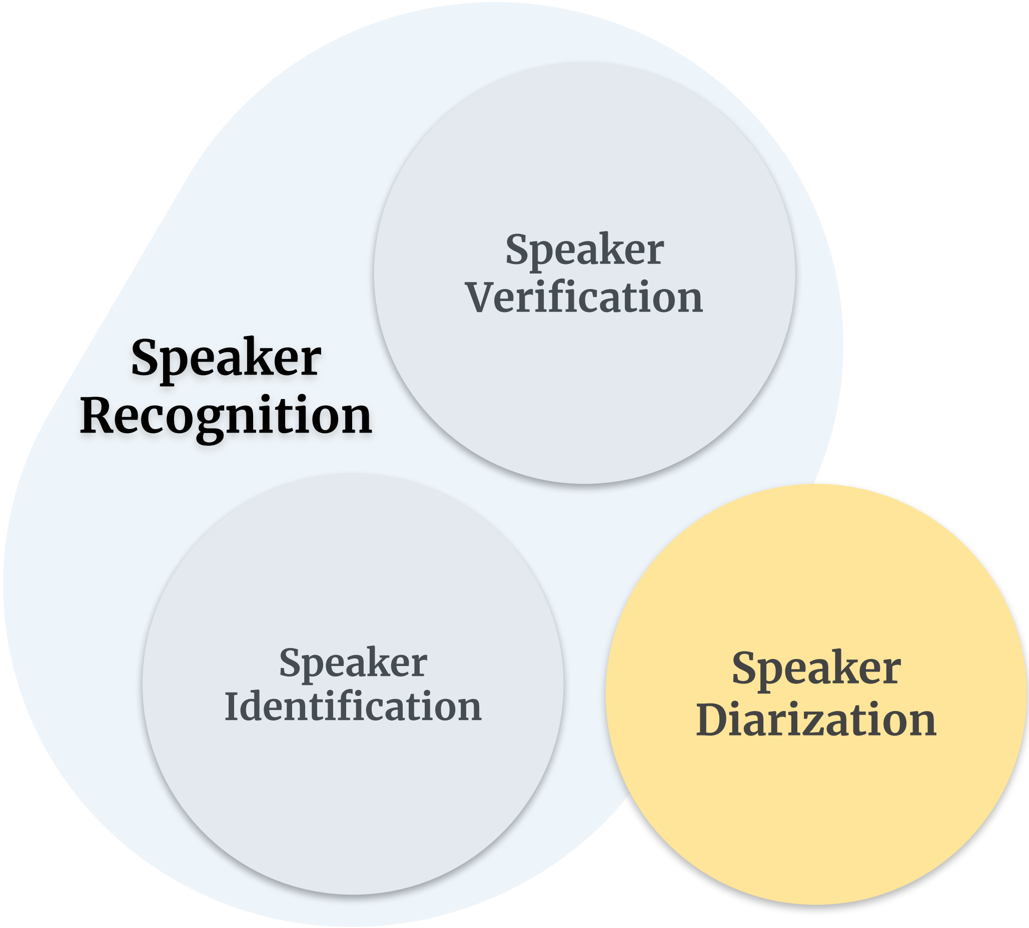
Introduction

Speaker Diarization



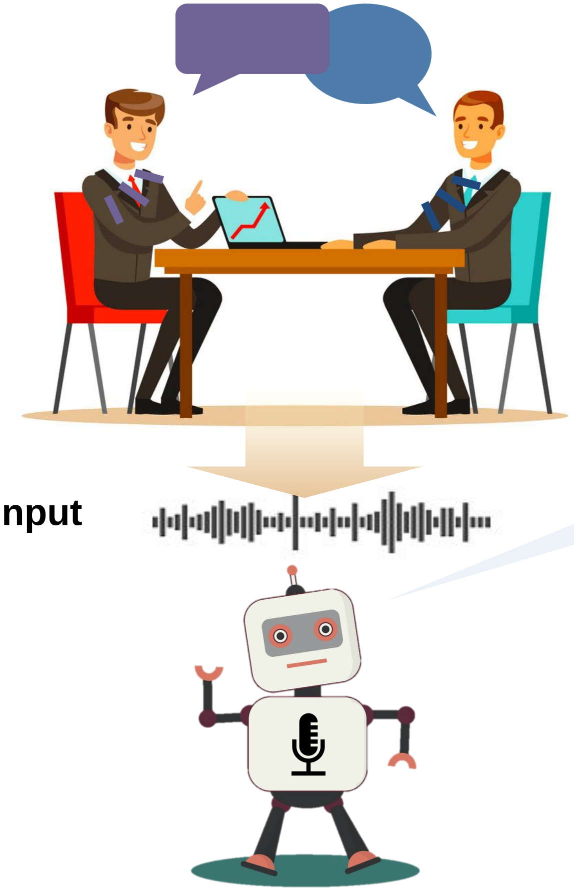
- Speaker diarization output = “Who spoke when?”
- Cluster the speech segments
- Does not identify each speaker

Introduction



Introduction

Why is Speaker Diarization Important ?



Audio Input

Automatic Speech Recognition
(ASR)

ASR Output ?

… how is your day going quite busy you must feel stressed out …

Speaker Diarization

Speaker A: how is your day going

Speaker B: quite busy

Speaker A: you must feel stressed out

Introduction

Why is speaker diarization important ?

Content managing and media indexing

Couple’s behavior study

Patient and caregiver

Meeting transcription

Introduction

Applications of Speaker Diarization

Where do we use speaker diarization for?

- Global pandemic led us to virtual world and created lots of applications for speaker diarization
- Lectures, interviews, office meetings and happy hours.
- The interactions between the participants need to be analyzed.

Katrin Kirchhoff (Amazon)

Introduction

Applications of Speaker Diarization

Where do we use speaker diarization for?

- Transcription for medical notes: Words and emotion
- Legal proceedings and court proceedings: Speaker information is very important
- Earnings calls: Announcements and QnA sessions. Very rapidly paced.
- Lectures: Lecturer and questions from the audience.

Douglas Reynolds (MIT Lincoln Lab)

Introduction

Applications of Speaker Diarization

Where can we use speaker diarization?

- Interviews and Conversations: Who is speaking during the conversation. (e.g. teacher student interactions)
- Online Videos (e.g. YouTube): Speaker diarization provides speaker information for video indexing.

Quan Wang (Google)

Introduction

Applications of Speaker Diarization

What would be the applications of diarization?

- Meetings: Who is speaking when
- Analytics on media: Indexing of speakers (speaker tracking)
- Political debates: Speaking time of each speaker
- Analysis of communications: Control towers in airports, Fearless Challenge by UT (Radio communications between astronauts and Huston)

Xavier Anguera (ELSA)

Chapter 1

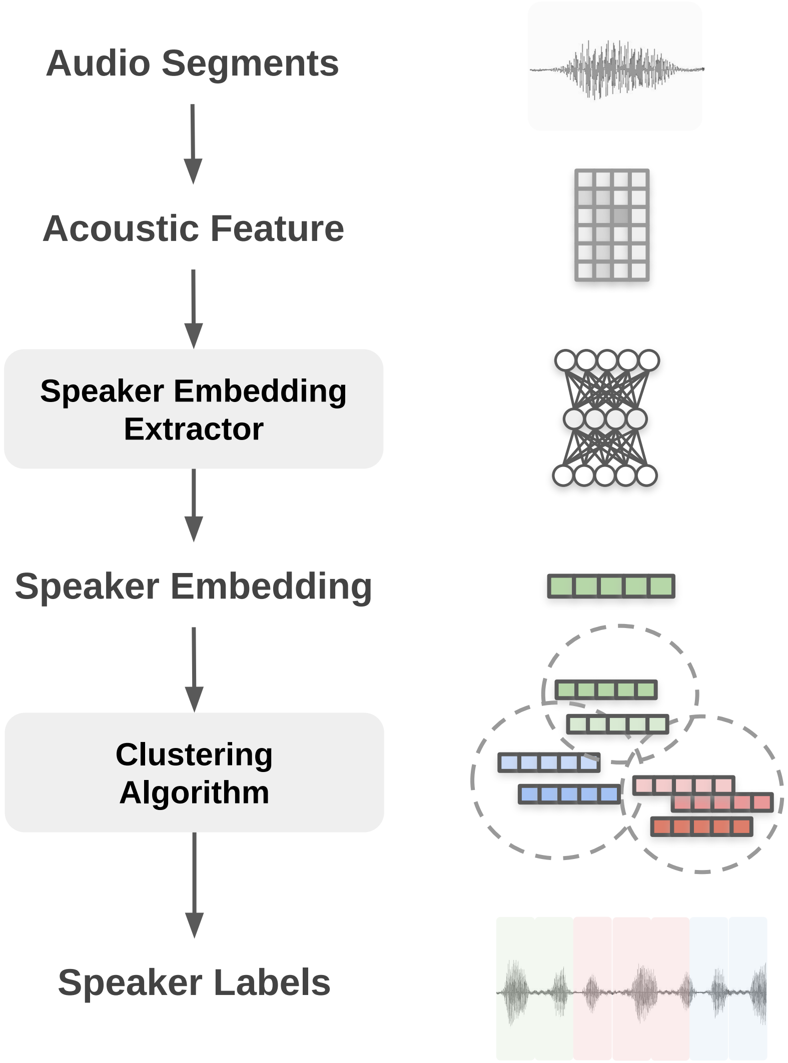
Diarization Overview

Part-2

Speaker Diarization Pipeline

Speaker Diarization Pipeline

Modular Diarization



Audio Segments

Acoustic Feature

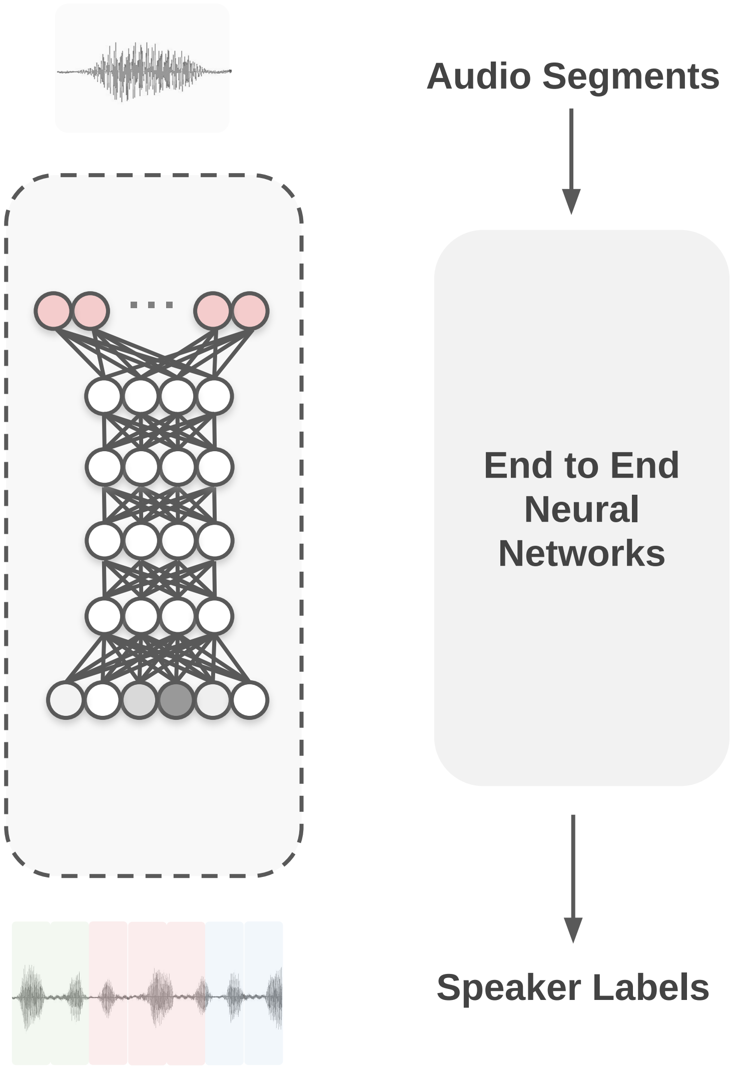
Speaker Embedding Extractor

Speaker Embedding

Clustering Algorithm

Speaker Labels

End2End Diarization



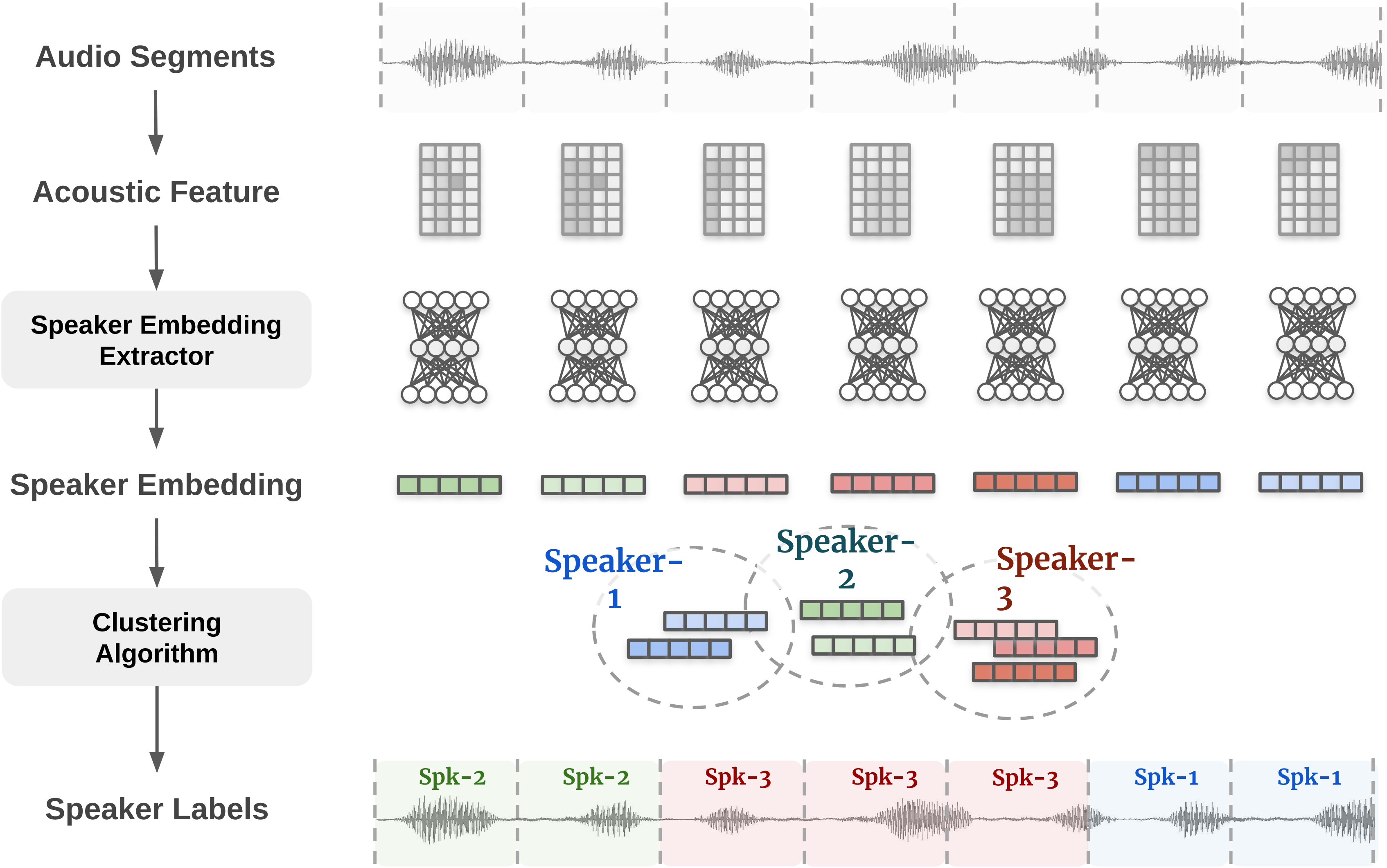
Audio Segments

End to End Neural Networks

Speaker Labels

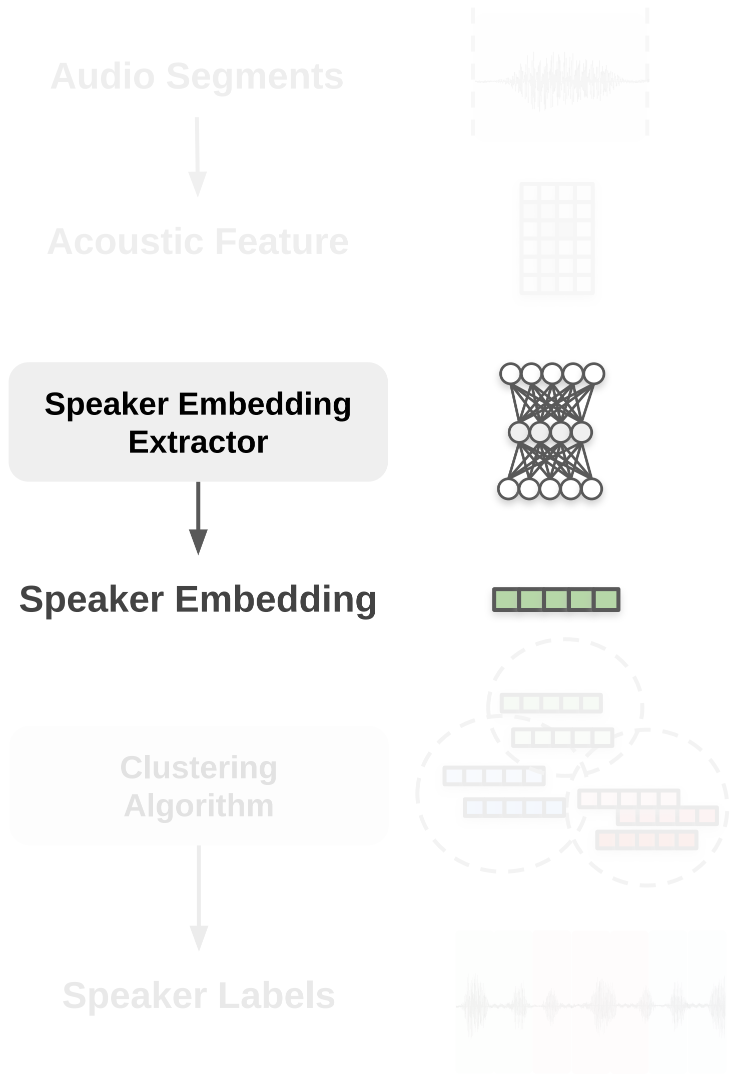
Speaker Diarization Pipeline

Speaker Diarization Pipeline



Speaker Diarization Pipeline

Speaker Diarization Pipeline: Speaker Representation



Audio Segments

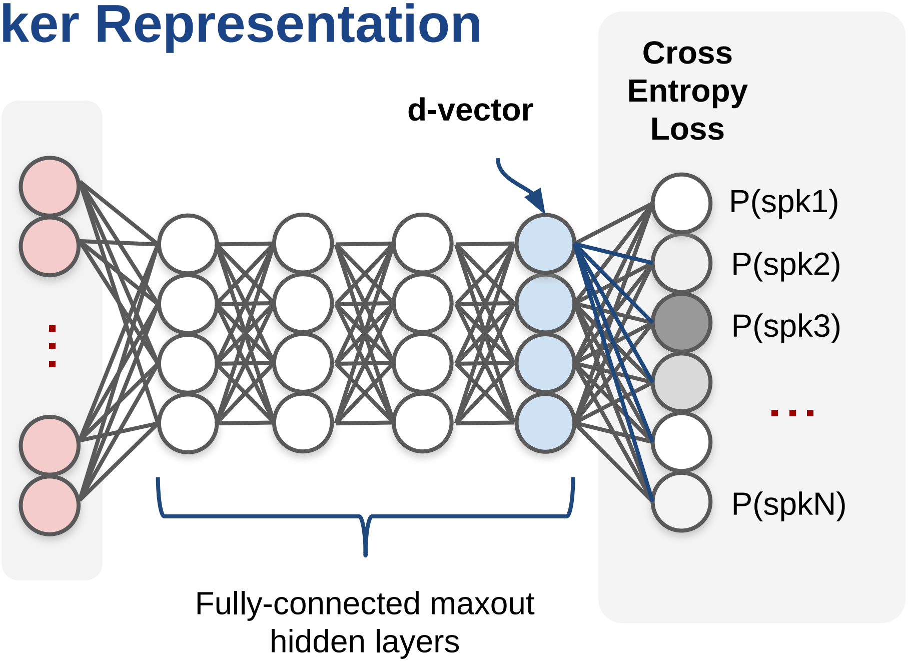
Acoustic Feature

Speaker Embedding Extractor

Speaker Embedding

Clustering Algorithm

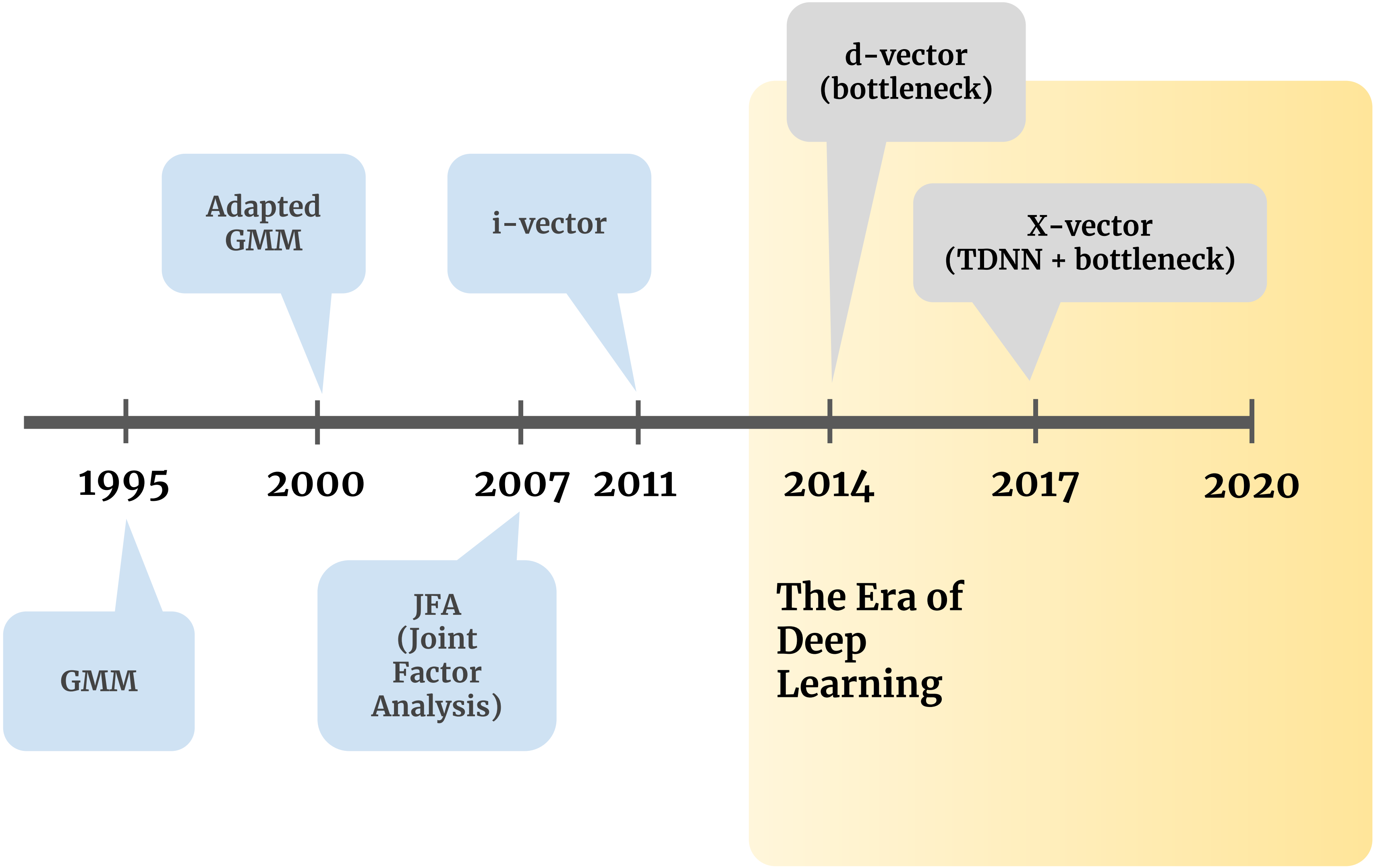
Speaker Labels



- Embedding is a dense vector of floating point numbers and represents the input (image, voice etc.)
- Embedding is also referred to as “representation”.
- Embedding is pulled from a bottleneck layer.
- [Example]

3.21	-1.25	-0.52	…	1.12	0.98	4.58

Speaker Diarization Pipeline



Adapted GMM

i-vector

d-vector (bottleneck)

X-vector (TDNN + bottleneck)

1995

2000

2007

2011

2014

2017

2020

GMM

JFA (Joint Factor Analysis)

The Era of Deep Learning

Speaker Diarization Pipeline

iVector representation

Fixed-length representation of speech utterances: Speaker characteristics into a floating point vector

$$M = m + Tw$$

$$p(\mathbf{x}|\lambda) = \mathcal{N}(\mathbf{x}|\mu, \Sigma)$$

(C□F ✕ 1) (C□F ✕ 1) (C□F ✕ D) (D ✕ 1)

(500 ✕ 1) (500 ✕ 1) (500 ✕ 400) (400 ✕ 1)

- **C**: Number of mixture components (25)
- **F**: Dimension of feature (20)

M

- Speaker and Channel **Dependent** GMM Supervector
- Normally distributed with *N(0,I)*
- Mean is ***m***, Covariance matrix is $\boldsymbol{T}\,\boldsymbol{T}^{t}$

m

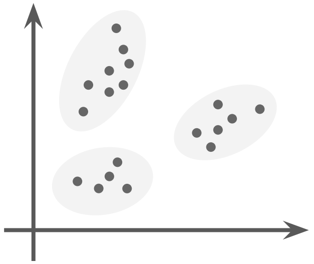
- Speaker and Channel **Independent** GMM Supervector
- Mean of ***M***

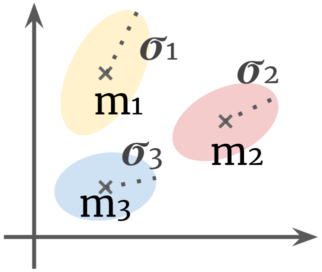
w

- **I-vector** (Identity vector)
- Speaker factor

T

- Total variability space
- Contains **speaker variability** and **session variability**
- Trained on data using Eigenvoice method





$$\mu = [m_1, m_2, \ldots, m_N] \quad \Sigma = \begin{pmatrix} \sigma_{1,1} & \ldots & \sigma_{1,N} \\ \ldots & \sigma_{i,j} & \ldots \\ \sigma_{N,1} & \ldots & \sigma_{N,N} \end{pmatrix}$$

$$N_c(m, \Sigma_c)$$

UBM (Universal Background Model)

Speaker Diarization Pipeline

i-vector representation

- **MAP (Maximum a Posterior) Estimation:**
- For this utterance y_t, what is the best i-vector to fit UBM model?

$$\widehat{\omega}_{MAP}(y) = \arg\max_{w} \ f(y \mid w) g(w) \qquad \Phi = \arg\max_{w} \left[\prod_{c=1}^{C} \prod_{t=1}^{N_c} N(y_t | m_c + T_c w, \Sigma_c) \right] N(w|0, \mathrm{I})$$

- **Solution of MAP estimator:**

$$N_c = \sum_{t=1}^{L} \mathrm{P}(c|y_t, \Omega) \quad \text{Constant} \qquad \tilde{F}_c = \sum_{t=1}^{L} \mathrm{P}(c|y_t, \Omega)(y_t - m_c) \quad \mathbf{(F \times 1)}$$

$$\widetilde{F}(u) = \begin{bmatrix} \tilde{F}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tilde{F}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \tilde{F}_C \end{bmatrix} \qquad N(u) = \begin{bmatrix} N_c\mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & N_c\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & N_c\mathbf{I} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Sigma_c \end{bmatrix}$$

$$w = (I + T^t \Sigma^{-1} N(u) T)^{-1} T^t \Sigma^{-1} \widetilde{F}(u)$$

- First Order Baum-Welch Statistics from sequence yt and UBM □ (c is component index)

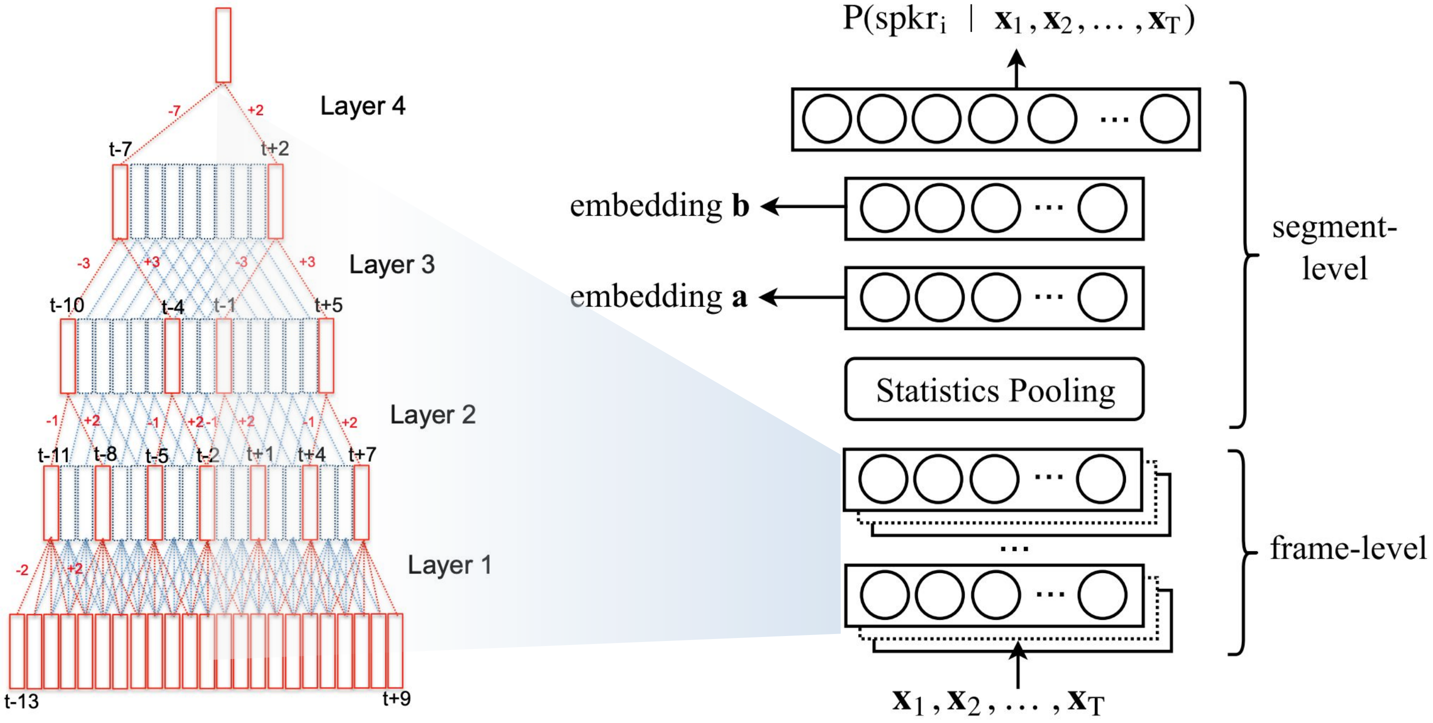
  $\tilde{F}(u)$: C xF1 Obtained **from each utterance** by using UBM □

  $\tilde{N}(u)$: CF xCF diagonal matrix whose diagonal blocks are $N_c I$

- **T** and **Σ** is obtained (trained) from training data using EM algorithm.
- D x 1 matrix as an output i-vector.
- Large inverse matrix → Time consuming Inference

Speaker Diarization Pipeline

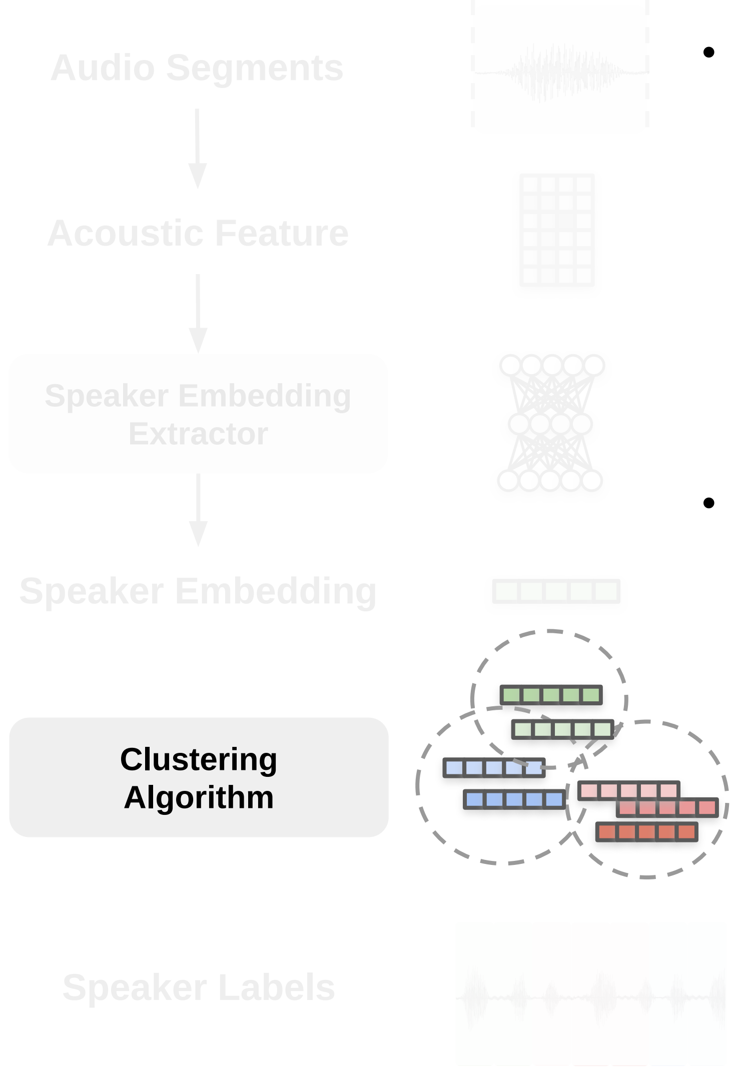
x-vector representation



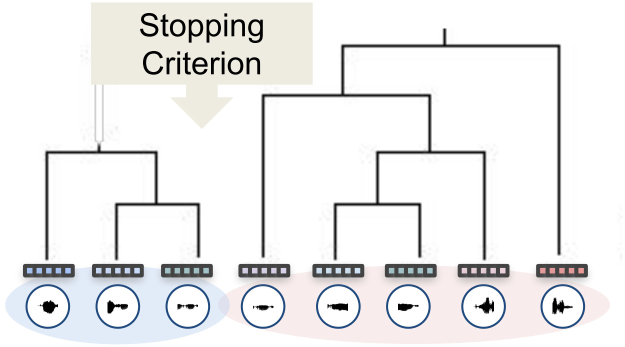
- 20 dimensional MFCC as input feature
- TDNN is basically 1-dimensional Convolutional Neural Networks
- Statistics pooling layer calculates mean and variance of final frame level layer
- 300 hidden units for a and 512 hidden units for b
- Cross-entropy loss function and PLDA scoring

Clustering and Speaker Counting

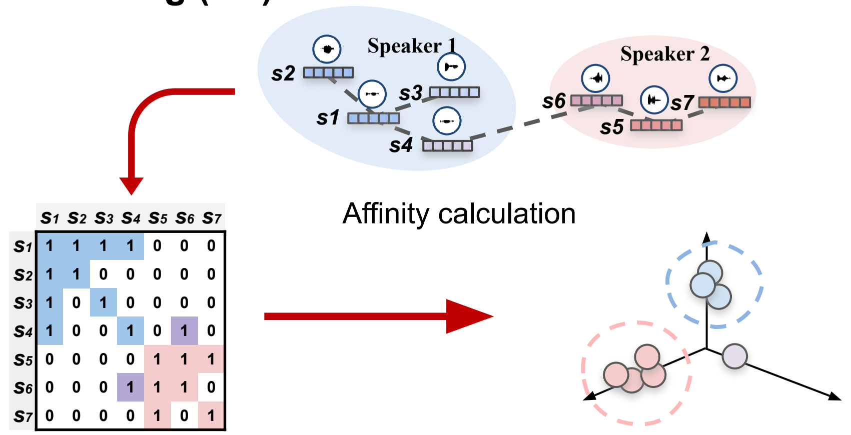
Speaker Diarization Pipeline: Clustering Method



- **Agglomerative Hierarchical Clustering (AHC)**



- **Spectral Clustering (SC)**



Affinity calculation

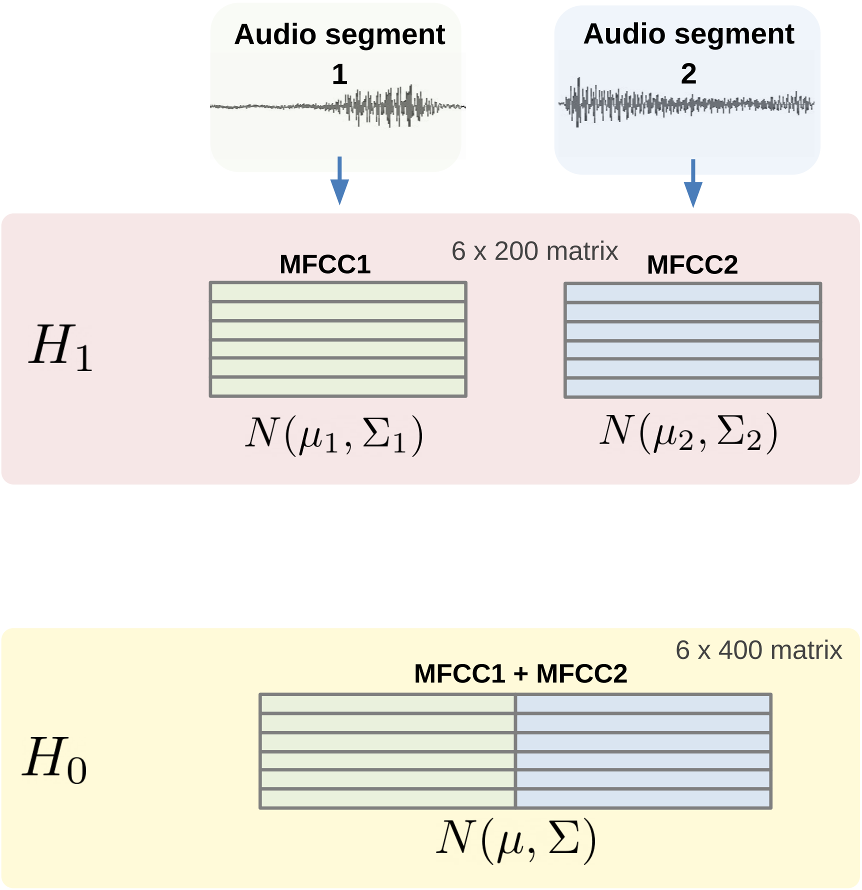
	s_1	s_2	s_3	s_4	s_5	s_6	s_7
s_1	1	1	1	1	0	0	0
s_2	1	1	0	0	0	0	0
s_3	1	0	1	0	0	0	0
s_4	1	0	0	1	0	1	0
s_5	0	0	0	0	1	1	1
s_6	0	0	0	1	1	1	0
s_7	0	0	0	0	1	0	1

Eigen value decomposition of affinity matrix

K-means of spectral data points

Clustering and Speaker Counting

Bayesian Information Criterion (BIC)



Audio segment 1

Audio segment 2

H_1

MFCC1

6 x 200 matrix

MFCC2

$N(\mu_1, \Sigma_1)$

$N(\mu_2, \Sigma_2)$

H_0

6 x 400 matrix

MFCC1 + MFCC2

$N(\mu, \Sigma)$

- Assume a Gaussian process

$$\boldsymbol{x}_i \sim N(\mu_i, \Sigma_i)$$

- Hypothesis testing

$$H_0 : x_1 \cdots x_N \sim N(\mu, \Sigma)$$

$$H_1 : x_1 \cdots x_i \sim N(\mu_1, \Sigma_1)$$

$$x_{i+1} \cdots x_N \sim N(\mu_2, \Sigma_2)$$

- Maximum likelihood ratio statistic:

$$R = \log\left(\frac{|\Sigma|^N}{|\Sigma_1|^{N_1}|\Sigma_2|^{N_2}}\right)$$

$$= N \log|\Sigma| - N_1 \log|\Sigma_1| - N_2 \log|\Sigma_2|$$

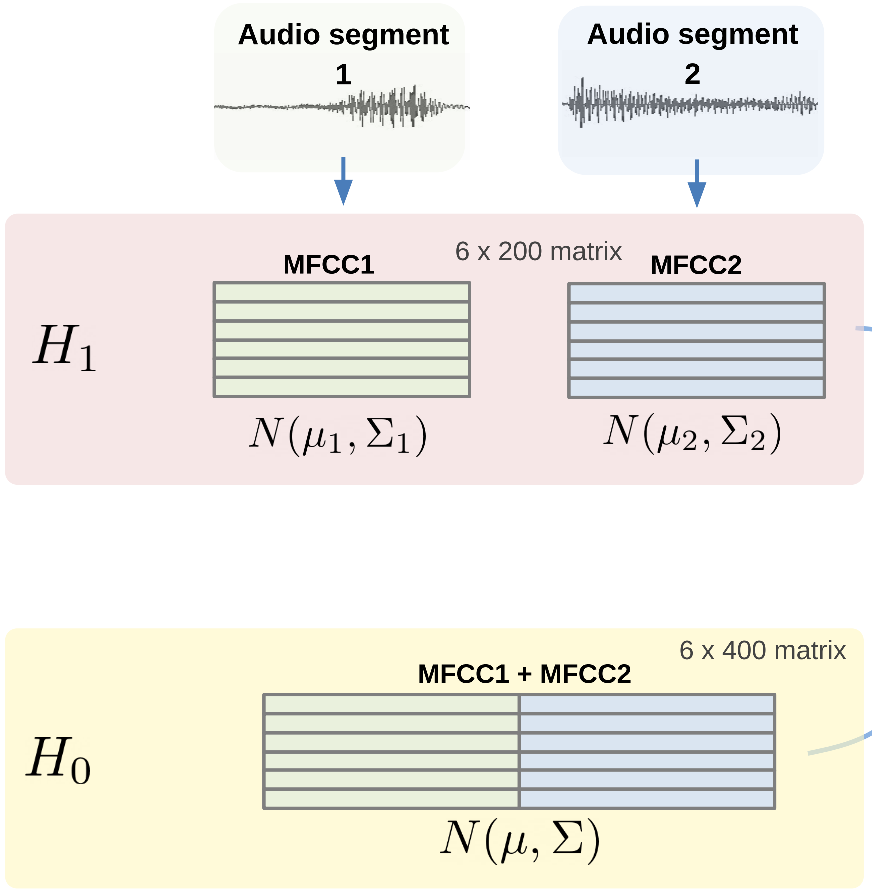
- BIC value

$$BIC = R - \lambda P$$

P : Dimensionality Compensation Factor

Clustering and Speaker Counting

Bayesian Information Criterion (BIC)



Audio segment 1

Audio segment 2

H_1

MFCC1

6 x 200 matrix

MFCC2

$N(\mu_1, \Sigma_1)$

$N(\mu_2, \Sigma_2)$

H_0

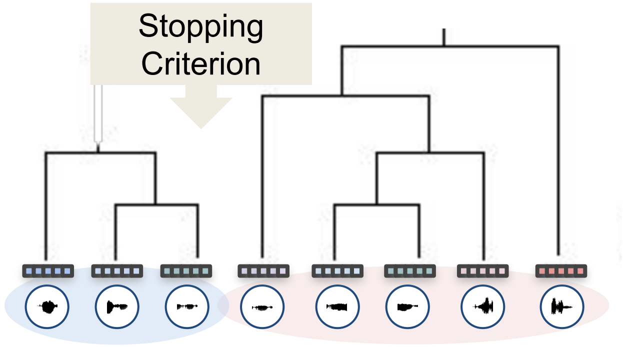
6 x 400 matrix

MFCC1 + MFCC2

$N(\mu, \Sigma)$

BIC-1 is from MFCC (Voice Characteristics)

$$BIC = R - \lambda P$$



Stopping Criterion

Clustering and Speaker Counting

Probabilistic Linear Discriminant Analysis (PLDA)

For i-th Speaker and j-th session:

Residual Variability

$$\phi_{ij} = \mu + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij} + \epsilon_{ij}$$

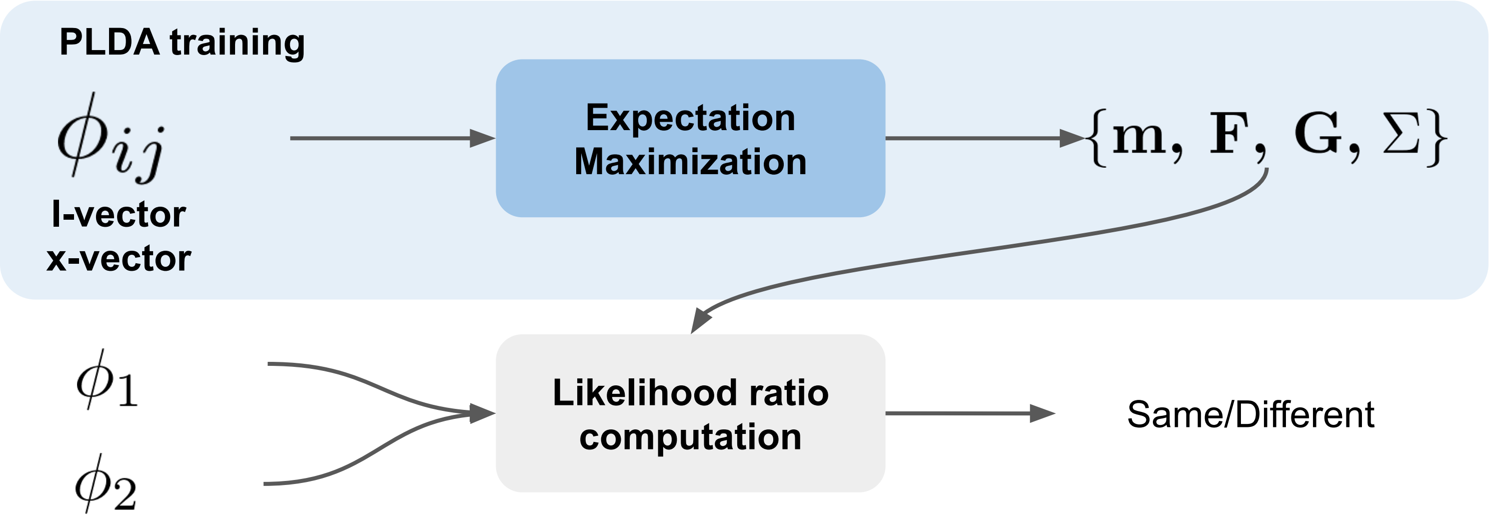
I-vector
x-vector

Speaker-dependent

Recording-dependent

$$= \mu + \mathbf{F}\mathbf{h}_i + \epsilon_{ij}$$

Simplified



PLDA training

ϕ_{ij}

I-vector
x-vector

Expectation Maximization

$\{\mathbf{m}, \mathbf{F}, \mathbf{G}, \Sigma\}$

ϕ_1

ϕ_2

Likelihood ratio computation

Same/Different

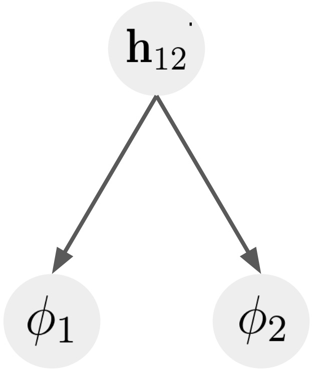
S.J.D. Prince, J.H. Elder, Probabilistic linear discriminant analysis for inferences about identity, in: Proceedings of International Conference on Computer Vision, 2007, pp. 1–8.
Rajan, P., Afanasyev, A., Hautamäki, V., & Kinnunen, T. (2014). From single to multiple enrollment i-vectors: Practical PLDA scoring variants for speaker verification. *Digital Signal Processing*, *31*

Clustering and Speaker Counting

Hypothesis H0: <u>Two samples are from the same speaker</u>

$$\begin{bmatrix}\phi_1\\ \phi_2\end{bmatrix} = \begin{bmatrix}\mu\\ \mu\end{bmatrix} + \left[\underbrace{\begin{matrix}\mathbf{F}\\ \mathbf{F}\end{matrix}}_{\mathbf{m}}\;\underbrace{\begin{matrix}\mathbf{G} & 0\\ 0 & \mathbf{G}\end{matrix}}_{\mathbf{A}}\right]\begin{bmatrix}\mathbf{h}_{12}\\ \mathbf{w}_1\\ \mathbf{w}_2\end{bmatrix} + \begin{bmatrix}\epsilon_1\\ \epsilon_2\end{bmatrix}$$

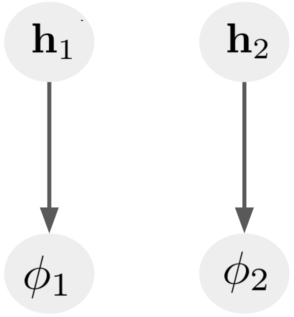
$$\log p(\phi_1, \phi_2 | H_0) = \log \mathcal{N}(\begin{bmatrix}\phi_1\\ \phi_2\end{bmatrix} | \, \mathbf{m}, \mathbf{A}\mathbf{A}^T + \Sigma)$$



Hypothesis H1: <u>Two samples are from different speakers</u>

$$\begin{bmatrix}\phi_1\\ \phi_2\end{bmatrix} = \begin{bmatrix}\mu\\ \mu\end{bmatrix} + \begin{bmatrix}\mathbf{F} & \mathbf{G} & 0 & 0\\ 0 & 0 & \mathbf{F} & \mathbf{G}\end{bmatrix}\begin{bmatrix}\mathbf{h}_1\\ \mathbf{w}_1\\ \mathbf{h}_2\\ \mathbf{w}_2\end{bmatrix} + \begin{bmatrix}\epsilon_1\\ \epsilon_2\end{bmatrix}$$

$$\log p(\phi_1, \phi_2 | H_1) = \sum_{l=1}^{2} \log \mathcal{N}(\phi_l | \, \mathbf{m}, \mathbf{F}\mathbf{F}^T + \mathbf{G}\mathbf{G}^T + \Sigma)$$



S.J.D. Prince, J.H. Elder, Probabilistic linear discriminant analysis for inferences about identity, in: Proceedings of International Conference on Computer Vision, 2007, pp. 1–8.
Rajan, P., Afanasyev, A., Hautamäki, V., & Kinnunen, T. (2014). From single to multiple enrollment i-vectors: Practical PLDA scoring variants for speaker verification. *Digital Signal Processing*, *31*

Clustering and Speaker Counting

Hypothesis H0 VS Hypothesis H1:

$$s(\phi_1, \phi_2) = \log p(\phi_1, \phi_2 | H_0) - \log p(\phi_1, \phi_2 | H_1)$$

$$s(\phi_1, \phi_2) = \frac{1}{2}(\psi_1^T + \psi_2^T)\mathbf{M}_2(\psi_1 + \psi_2) - \frac{1}{2}\psi_1^T \mathbf{M}_2 \psi_1 - \frac{1}{2}\psi_2^T \mathbf{M}_2 \psi_2 + K$$

$$\mathbf{M}_J = [J\mathbf{F}^T[\mathbf{G}\mathbf{G}^T + \Sigma]^{-1}\mathbf{F} + \mathbf{I}]^{-1}$$

$$K = \frac{1}{2}\log|\mathbf{M}_2| - \log|\mathbf{M}_1|$$

Constant for given set of parameters

$$\psi_k = \mathbf{F}^T[\mathbf{G}\mathbf{G}^T + \Sigma]^{-1}(\phi_l - \mathbf{m})$$

- ψ variable centralizes the input i-vector(ϕ)
- Projects it onto the subspace **F** to co-vary the most
- de-emphasizing the subspace **G** pertaining to channel variability.
- Ideally, stopping criterion should be 0, but **in practice it varies from -0.5~0.5** and needs to be tuned on development set.

S.J.D. Prince, J.H. Elder, Probabilistic linear discriminant analysis for inferences about identity, in: Proceedings of International Conference on Computer Vision, 2007, pp. 1–8.
Rajan, P., Afanasyev, A., Hautamäki, V., & Kinnunen, T. (2014). From single to multiple enrollment i-vectors: Practical PLDA scoring variants for speaker verification. *Digital Signal Processing*, *31*

Clustering and Speaker Counting

Speaker Counting is Hard!

- For example, meetings with more than 10 speakers could be very challenging
- Not that many studies have been done for estimating a large number of speakers.
- Large meetings and cocktail parties remain as challenging scenarios for speaker diarization.

Katrin Kirchhoff (Amazon)

Clustering and Speaker Counting

Speaker counting in real life scenarios

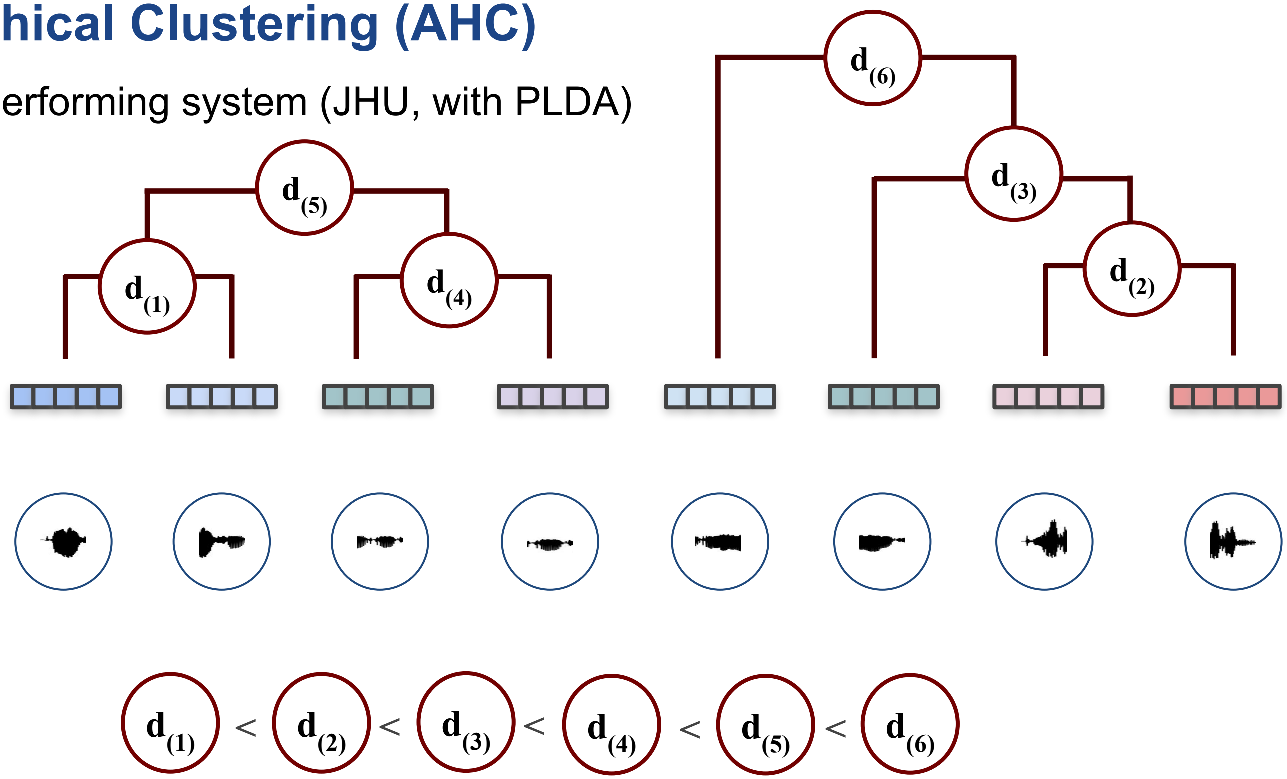
- Large number of speakers makes diarization very challenging.
- Providing the number of speakers to the diarization system can be advantageous.

Gakuto Kurata (IBM)

Clustering and Speaker Counting

Agglomerative Hierarchical Clustering (AHC)

Appeared in DIHARD-I: Best performing system (JHU, with PLDA)



Speaker Representation

Audio Segments

Average Distance in order

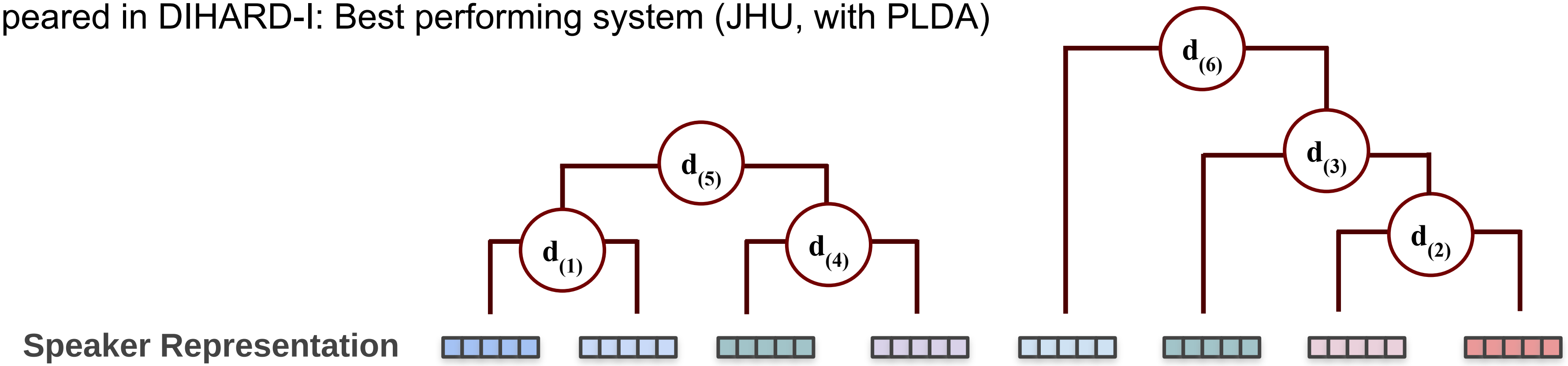
$d_{(1)} < d_{(2)} < d_{(3)} < d_{(4)} < d_{(5)} < d_{(6)}$

- Merge the closest pairs based on a specific distance measure **d**.
- We can either stop at:
  - When the number of clusters are reduced to **N-clusters**
  - When the shortest distance among clusters reaches stopping threshold **dc**.
    (**dc** needs supervised tuning)

Clustering and Speaker Counting

Agglomerative Hierarchical Clustering (AHC)

Appeared in DIHARD-I: Best performing system (JHU, with PLDA)



BIC $R(i) = \log\left(\frac{|\Sigma|^N}{|\Sigma_1|^{N_1}|\Sigma_2|^{N_2}}\right) - \lambda P$ $= 0$

Cosine Similarity $cos(\boldsymbol{e}_1, \boldsymbol{e}_2) = \frac{\boldsymbol{e}_1 \cdot \boldsymbol{e}_2}{||\boldsymbol{e}_1|| \cdot ||\boldsymbol{e}_2||}$ $= d_c$

Stopping Criterion

PLDA $s(\phi_1, \phi_2) = \log p(\phi_1, \phi_2|H_0) - \log p(\phi_1, \phi_2|H_1)$ $= 0$ or $= d_c$

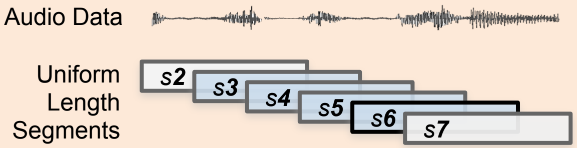
- For AHC, stopping criterion should be optimized on development dataset.
- In theory, the threshold should be 0 for PLDA distance measure, but in practice, the ideal threshold varies from -0.5 to 0.5
- In AHC approach, speaker counting can be very dependent on the stopping threshold.

Clustering and Speaker Counting

Spectral Clustering (SC) with binarized cosine similarity

Appeared in CHIME-6 track 2 Challenge Winning System (STC)

Audio Segmentation



Binarization

$$\mathbf{A}_p = binarize(\mathbf{A}, p)$$

Symmetrization

$$\bar{\mathbf{A}}_p = \frac{1}{2}(\mathbf{A}_p + \mathbf{A}_p^T)$$

	s1	*s2*	*s3*	*s4*	*s5*	*s6*	*s7*	*s8*
s1	1	1	0	0	0	0	1	1
s2	1	1	0	0.5	0	0	1	0.5
s3	0	0	1	0	1	0.5	0.5	0
s4	0	0.5	0	1	0.5	1	0	0
s5	0	0	1	0.5	1	1	0.5	0
s6	0	0	0.5	1	1	1	0.5	0.5
s7	1	1	0.5	0	0.5	0.5	1	1
s8	1	0.5	0	0	0	0.5	1	1

Unnormalized Laplacian matrix

$$\begin{aligned} d_i &= \textstyle\sum_{k=1}^{N} a_{ik} \\ \mathbf{D}_p &= \mathrm{diag}\{d_1, d_2, ..., d_N\} \\ \mathbf{L}_p &= \mathbf{D}_p - \bar{\mathbf{A}}_p, \end{aligned}$$

Eigen Value Decomposition (EVD)

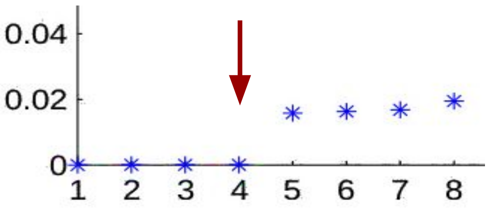
$$\mathbf{L}_p = \mathbf{U}_p \mathbf{\Sigma}_p \mathbf{V}_p^T$$

Eigengap Vector

$$\mathbf{e}_p = [\lambda_{p,2} - \lambda_{p,1}, \cdots, \lambda_{p,N} - \lambda_{p,N-1}]$$

Eigenvalues and eigengaps

Maximum Gap:
Number of Speakers

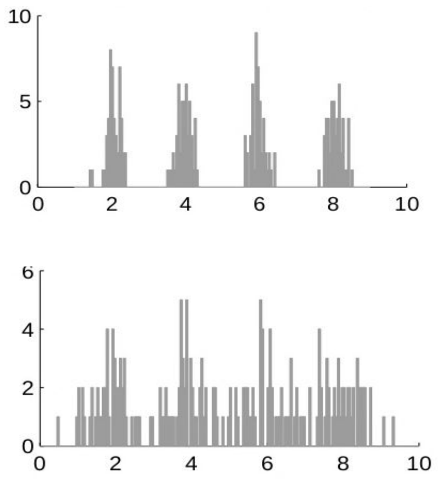


*Taejin Park et. al. “Auto-Tuning Spectral Clustering for Speaker Diarization Using Normalized Maximum Eigengap” IEEE SPL. 2019, p.381-385.

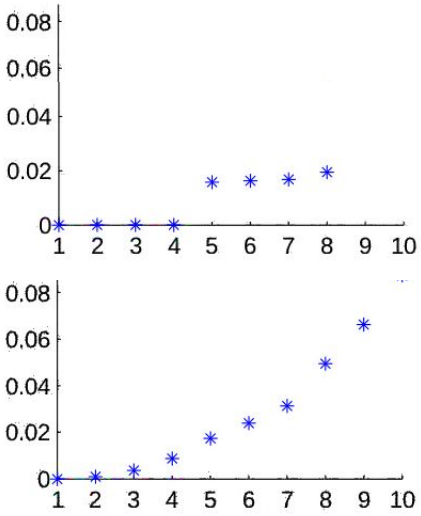
Clustering and Speaker Counting

Spectral Clustering (SC)

- Eigenvalues



- Eigenvalues and eigengaps



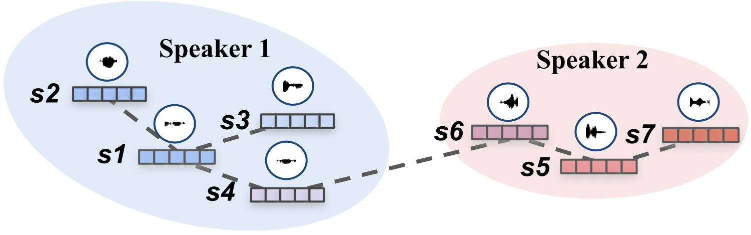
- Number of speakers can be estimated by the **maximum eigengap**.
- **Benefit:**
  - Eigengap based speaker number estimation is **less dependent on clustering parameter**.
- **Downside:**
  - Cannot compute huge session which will make a huge affinity matrix.
  - Spectral clustering and eigengap approach **can hardly be online fashion**.

[1] Taejin Park, Kyu Han, Manoj Kumar and Shrikanth Narayanan, “Auto-Tuning Spectral Clustering for Speaker Diarization Using Normalized Maximum Eigengap” IEEE Signal Processing Letters. 2019, p.381-385.

Clustering and Speaker Counting

Spectral Clustering (SC) with binarized cosine similarity

- Does not need PLDA, works with simple cosine similarity.
- High complexity but speaker counting performs better over PLDA+AHC



Affinity calculation $\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$

affinity matrix

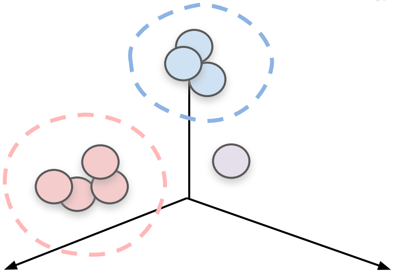
	s_1	s_2	s_3	s_4	s_5	s_6	s_7
s_1	1	1	1	1	0	0	0
s_2	1	1	0	0	0	0	0
s_3	1	0	1	0	0	0	0
s_4	1	0	0	1	0	1	0
s_5	0	0	0	0	1	1	1
s_6	0	0	0	1	1	1	0
s_7	0	0	0	0	1	0	1

Unnormalized Laplacian matrix

$$\begin{aligned} d_i &= \sum_{k=1}^{N} a_{ik} \\ \mathbf{D}_p &= \text{diag}\{d_1, d_2, ..., d_N\} \\ \mathbf{L}_p &= \mathbf{D}_p - \bar{\mathbf{A}}_p, \end{aligned}$$

Spectral Embeddings

$$\mathbf{L}_p = \mathbf{U}_p \mathbf{\Sigma}_p \mathbf{V}_p^T$$



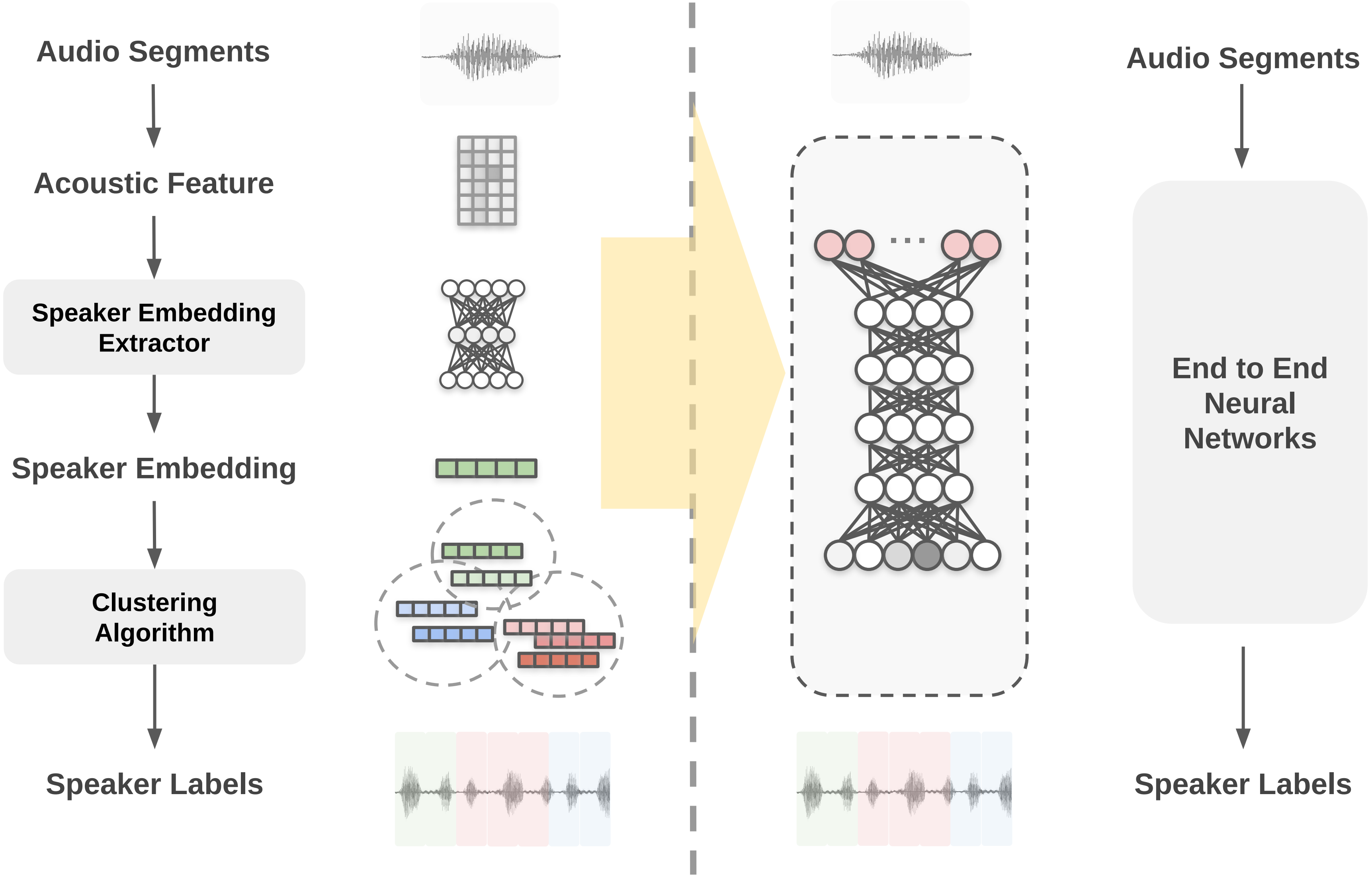
K-means of eigenvectors

Taejin Park, Kyu Han, Manoj Kumar and Shrikanth Narayanan, "Auto-Tuning Spectral Clustering for Speaker Diarization Using Normalized Maximum Eigengap" IEEE Signal Processing Letters. 2019, p.381-385.

Modular System vs End-to-End System

Modular Diarization

End2End Diarization



Audio Segments

Acoustic Feature

Speaker Embedding Extractor

Speaker Embedding

Clustering Algorithm

Speaker Labels

Audio Segments

End to End Neural Networks

Speaker Labels

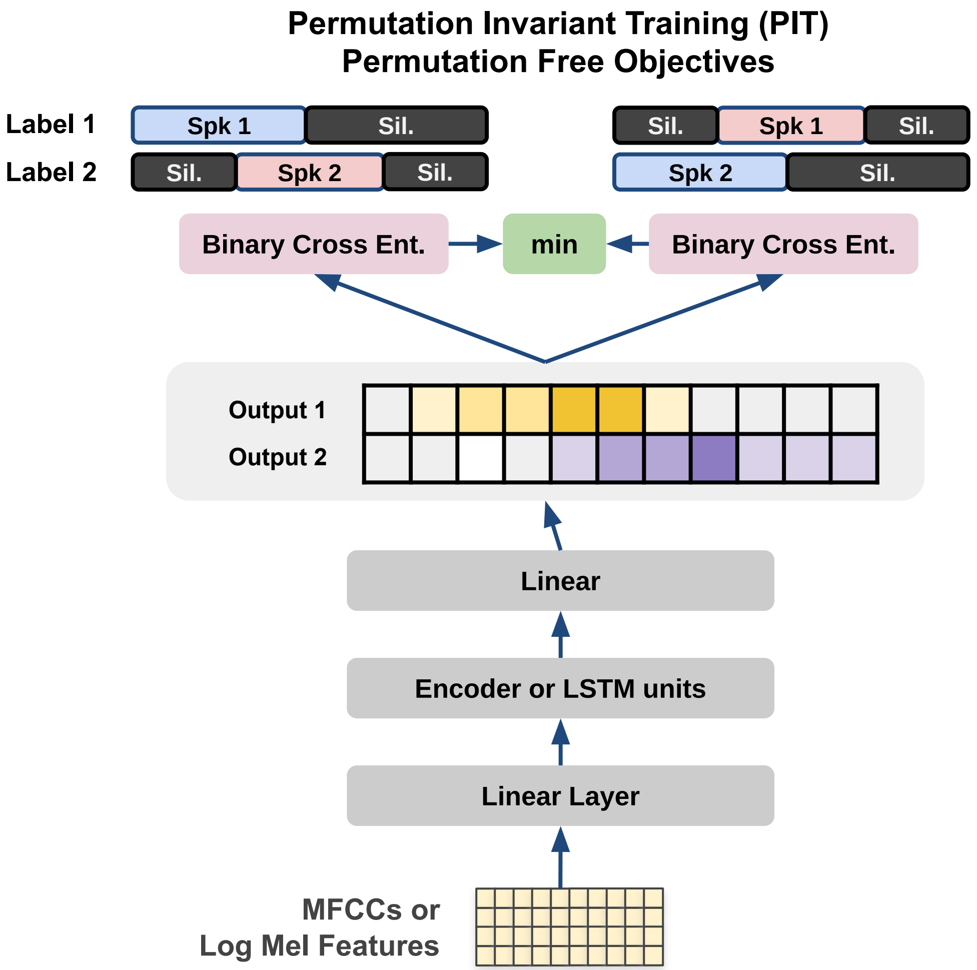
Modular System vs End-to-End System

End2End Diarization with Permutation invariant training

- A neural network model that accepts speech input and outputs speaker labels.
- End-to-end speaker diarization is not a downstream task
- Special type of loss calculation method is needed (e.g. PIT)

Benefits:

- Easy to train and deploy the model.
- Online-friendly architecture
- Fast inference speed



Fujita, Yusuke, et al. "End-to-end neural speaker diarization with permutation-free objectives." arXiv preprint arXiv:1909.05952 (2019).

Modular System vs End-to-End System

Modular Diarization VS End2End Diarization

	Modular Diarization	End2End Diarization
SoTA (Oct 2020) on CallHome Dataset	[1]Spk. Err 5~6% (System SAD) [1]DER 6~7% (Oracle SAD)	[2]Spk. Err > 10% (System SAD)
Training Data	Relatively **easy** to get (Separately train each module: embedding, clustering, language model)	Relatively **hard** to get balanced data Number of speakers Acoustic environment Language
Training Steps	Relatively **complicated**	Relatively **simple**
Validation of Each Function	Relatively **easy** (Separately test segmentation, embedding and clustering)	Relatively **hard**
Proper Applications	Media indexing Offline dialogue analysis	Online ASR pipeline Real-time dialogue system

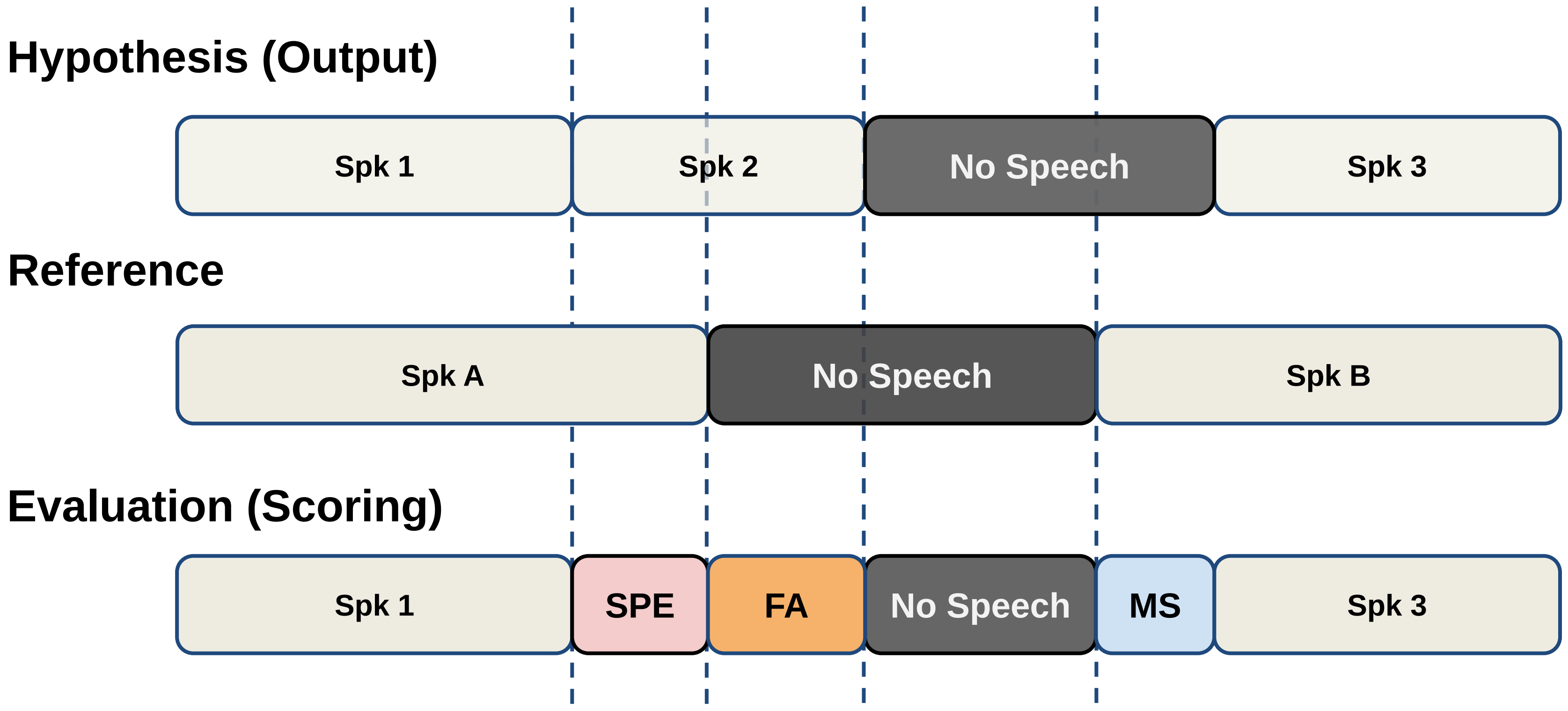
[1] **Fujita, Yusuke, et al. "End-to-End Neural Speaker Diarization with Self-attention."** ***arXiv preprint arXiv:1909.06247*****, 2019**

[2] **Lin, Qingjian, et al. "LSTM based Similarity Measurement with Spectral Clustering for Speaker Diarization." Interspeech 2019**

Diarization Evaluation

Traditional Diarization Error Rate (DER) – System SAD

How do we measure the accuracy of diarization? – With real-life SAD



Hypothesis (Output)

Spk 1 | Spk 2 | No Speech | Spk 3

Reference

Spk A | No Speech | Spk B

Evaluation (Scoring)

Spk 1 | SPE | FA | No Speech | MS | Spk 3

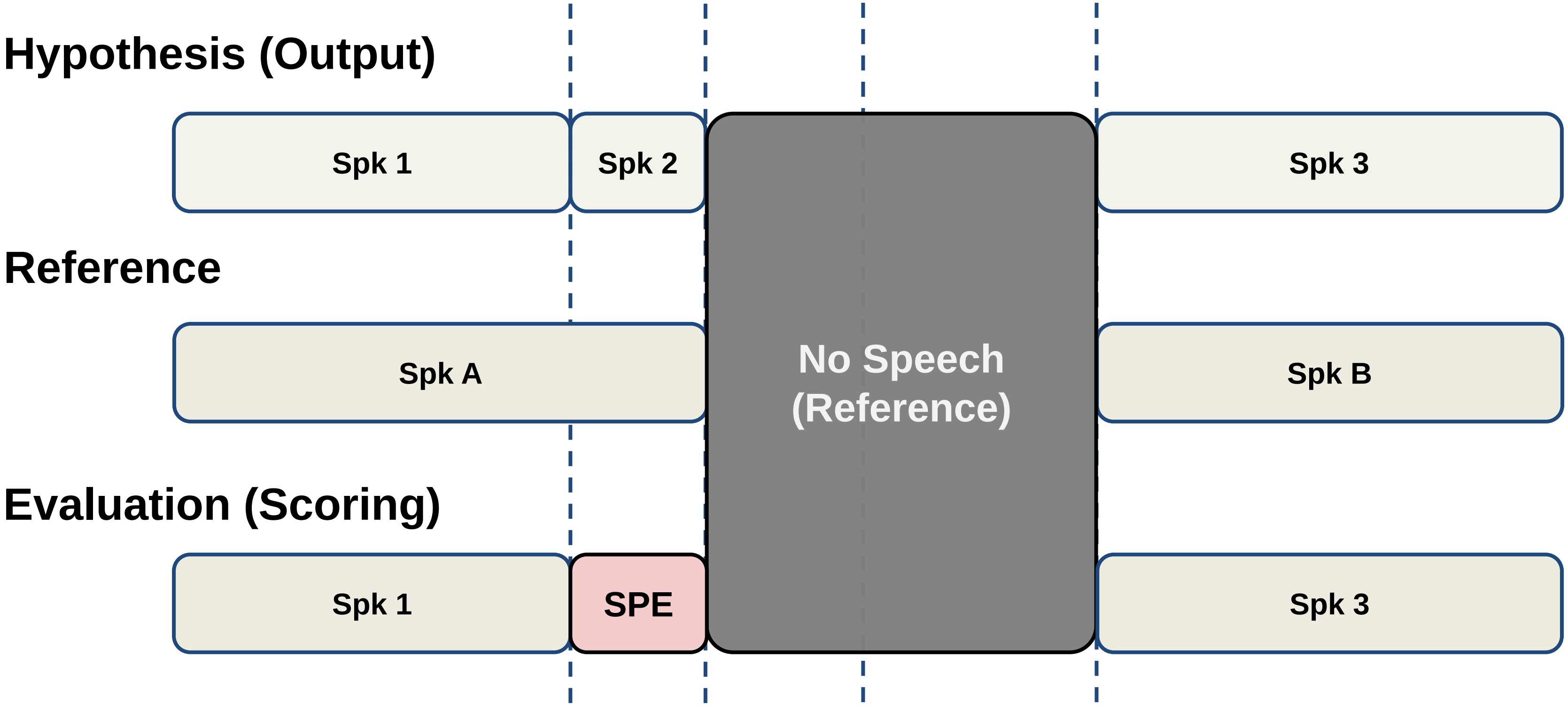
(% of time)

$$\text{DER} = \frac{\text{Speaker Error (SPE)} + \text{False Alarm (FA)} + \text{Missed Speech (MS)}}{\text{TOTAL}}$$

Diarization Evaluation

Traditional Diarization Error Rate (DER) – Oracle SAD

- With oracle speech activity detection time stamps



(% of time)

DER = Speaker Error (SPE)

- Speech activity information is given (always correct).
- Factors out the contribution of system SAD.

Diarization Evaluation

Jaccard Error Rate (JER)

Motivation for Jaccard Error Rate (JER)

- DER is biased towards the dominant speaker.
- Inactive speaker problem: a speaker that only appears for 10% of dialogue
- Alternative method is needed to address this problem.

Sriram Ganapathy (IISC)

Diarization Evaluation

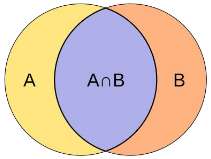
Jaccard Error Rate (JER)

- **“All speakers should be evaluated equally”**

$$\mathrm{JER}_{ref} = \frac{\mathrm{FA} + \mathrm{MISS}}{\mathrm{TOTAL}}$$

For a speaker

cf.) Jaccard Index



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

- **FA** is the total system speaker time **NOT** attributed to the reference speaker.
- **MISS** is the total reference speaker time **NOT** attributed to the system speaker
- **TOTAL**: The duration of the **union of reference and system speaker segments**

- **After Speaker matching between system output and reference (with no weights):**

$$\mathrm{JER} = \frac{1}{N} \sum_{ref} \mathrm{JER}_{ref}$$

- **JER and DER are highly correlated**
  - with JER typically being higher
  - Especially in recordings where one or more speakers is particularly dominant.

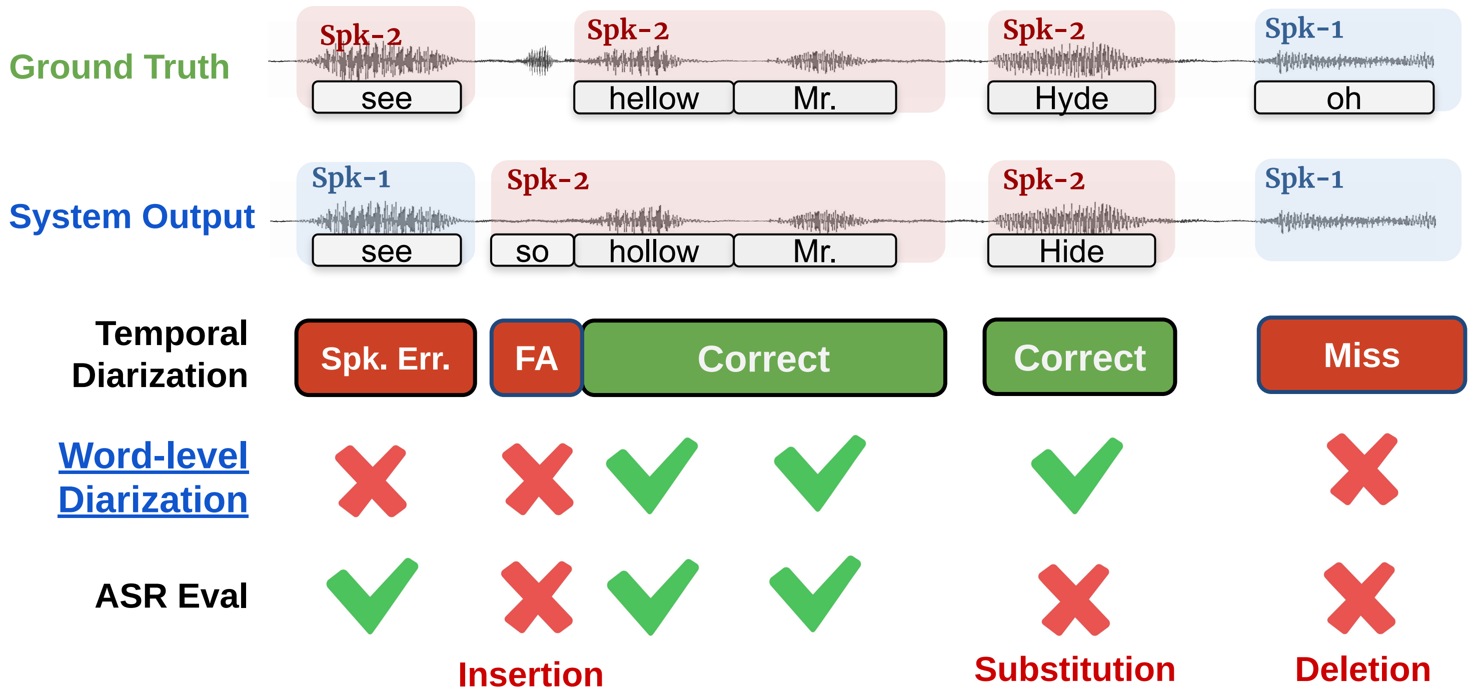
- **Where DER can easily exceed 500%, JER will never exceed 100%**

Ryant, Neville, et al. "Third DIHARD Challenge Evaluation Plan." *arXiv preprint arXiv:2006.05815* (2020).

Diarization Evaluation

Word Diarization Error Rate (WDER)

DER is not practical since diarization output is mostly displayed with words.



$$\text{WDER} = \frac{\text{Insertion (FA)} + \text{Deletion (Miss)} + \text{Speaker confusion} + \text{Word Substitution}^{*}}{\text{Number of Words in Reference}}$$

There are multiple versions of WDER depending on the numerator.

Diarization Evaluation

Evaluation Metric: WDER

Is Word-level DER useful?

- Rev.ai is positive on WDER and has internal measure, DER1, that is similar to WDER.
- In practice, diarization output is always accompanied by words.
- One drawback is: WDER has to be used with ASR WER because of deletion and insertion.
- We believe that WDER could be a good indication.

Miguel Jette (Rev.ai)

Chapter 1

Diarization Overview

Part-3

The Future of Speaker Diarization

The Future of Speaker Diarization

How far have we reached?

Traditional
Speaker Diarization Systems

Human Listeners

- **Supervised tuning is required**
  - Segmentation, embedding and clustering
- **Only use single modality (audio)**
  - Acoustic features to embedding
- **No contextual information is involved**
  - Easily fails when audio feature degrades

- **Require less of explicit tuning**
  - Humans do not learn the task separately:
  - Humans act more like End-to-end system (Simultaneously optimized)
- **Exploit many different modalities**
  - Lexical context, role recognition etc.
- **Consider contextual information**
  - Very robust even if one modality degrades (ex. What if identical twins talk?)

The Future of Speaker Diarization

The next generation diarization:
What will be discussed in the following chapters ?

- **Modularized to End-to-End System**
  - End-to-end system is easy to train and deploy
  - End-to-end system has straight-forward optimization process.
  - Good amount of training is needed to obtain a decent performance
- **Contextual Input : Speech Recognition with Diarization**
  - Word stream from ASR that provides **contextual information** for diarization.
  - Lexical input can be leveraged for improving speaker diarization
  - Joint training of speaker diarization **and** ASR + etc.
- **In the wild speaker diarization**
  - Overlap, short-segment speech
  - Domain mismatch
  - Inference Speed
  - Online Diarization
  - Training data for end-to-end system

Chapter 2

Speaker Diarization
and
Automatic Speech Recognition

Speaker Diarization and Automatic Speech Recognition

1. **Part 1: Speaker diarization enhanced by ASR outputs**

    1.1. Rich Transcription
    1.2. Diarization error rate (DER) vs word error rate (WER)
    1.3. Word boundaries from ASR for speaker diarization
    1.4. Speaker names in broadcast news

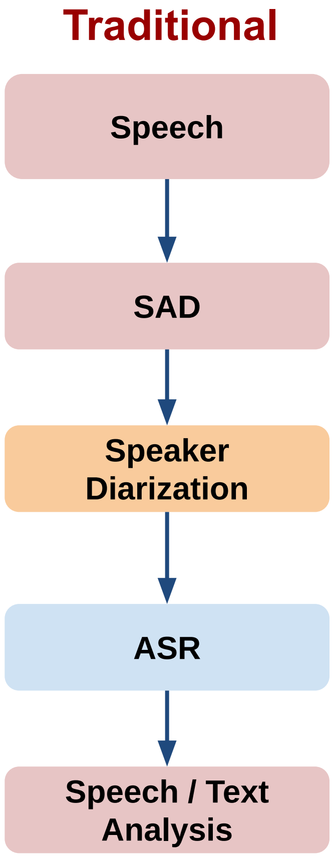
2. **Part 2: Lexical information used in speaker diarization**

    2.1. Segmentation using acoustic + lexical information
    2.2. Spectral clustering using acoustic + lexical information

3. **Part 3: Joint modeling of speaker diarization and ASR**

    3.1. Joint modeling of speaker diarization and ASR via sequence transduction
    3.2. Speaker diarization in target-speaker (TS) ASR
    3.3. SpeakerBeam

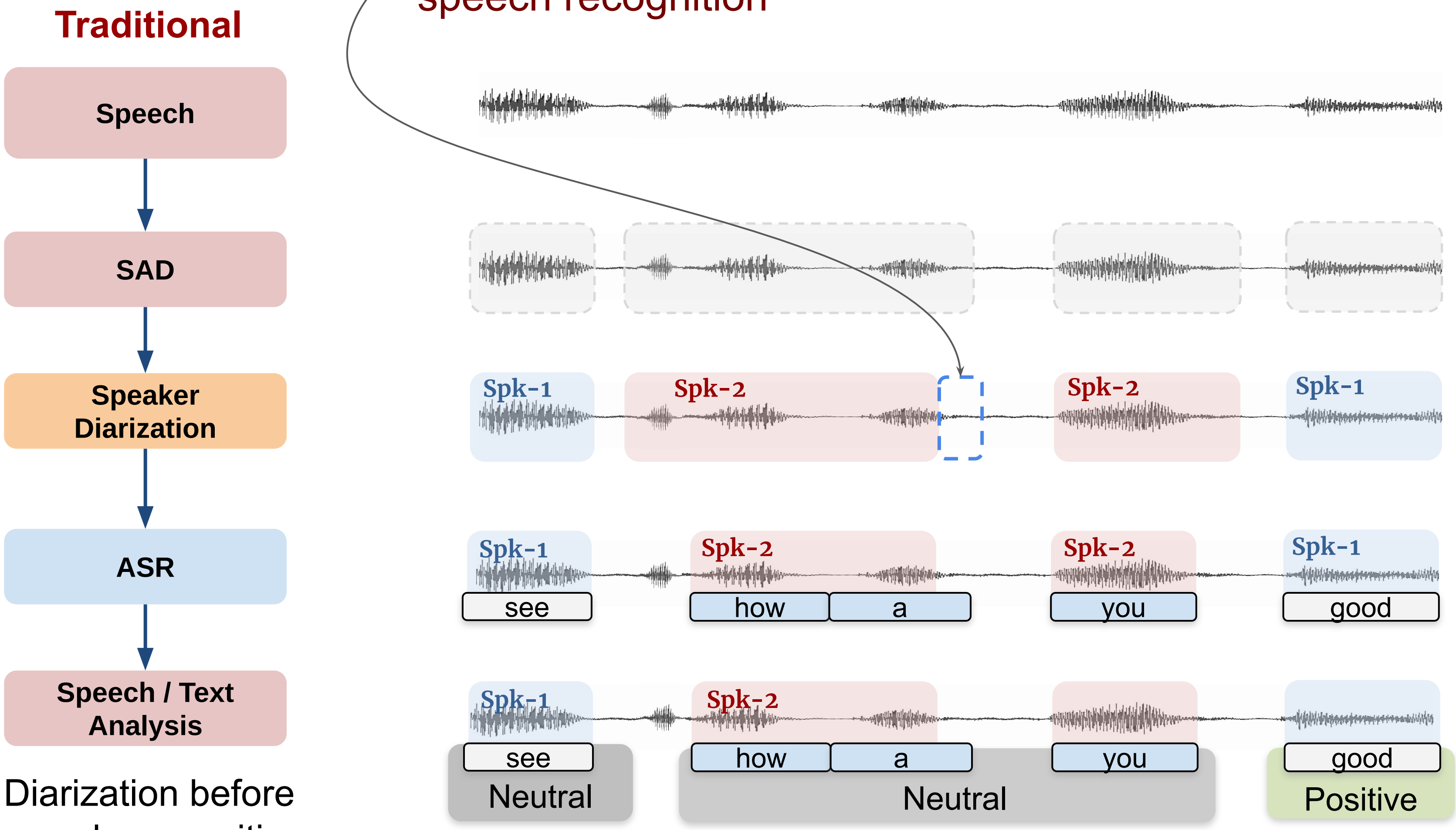
Speaker Diarization and Automatic Speech Recognition



Diarization before speech recognition

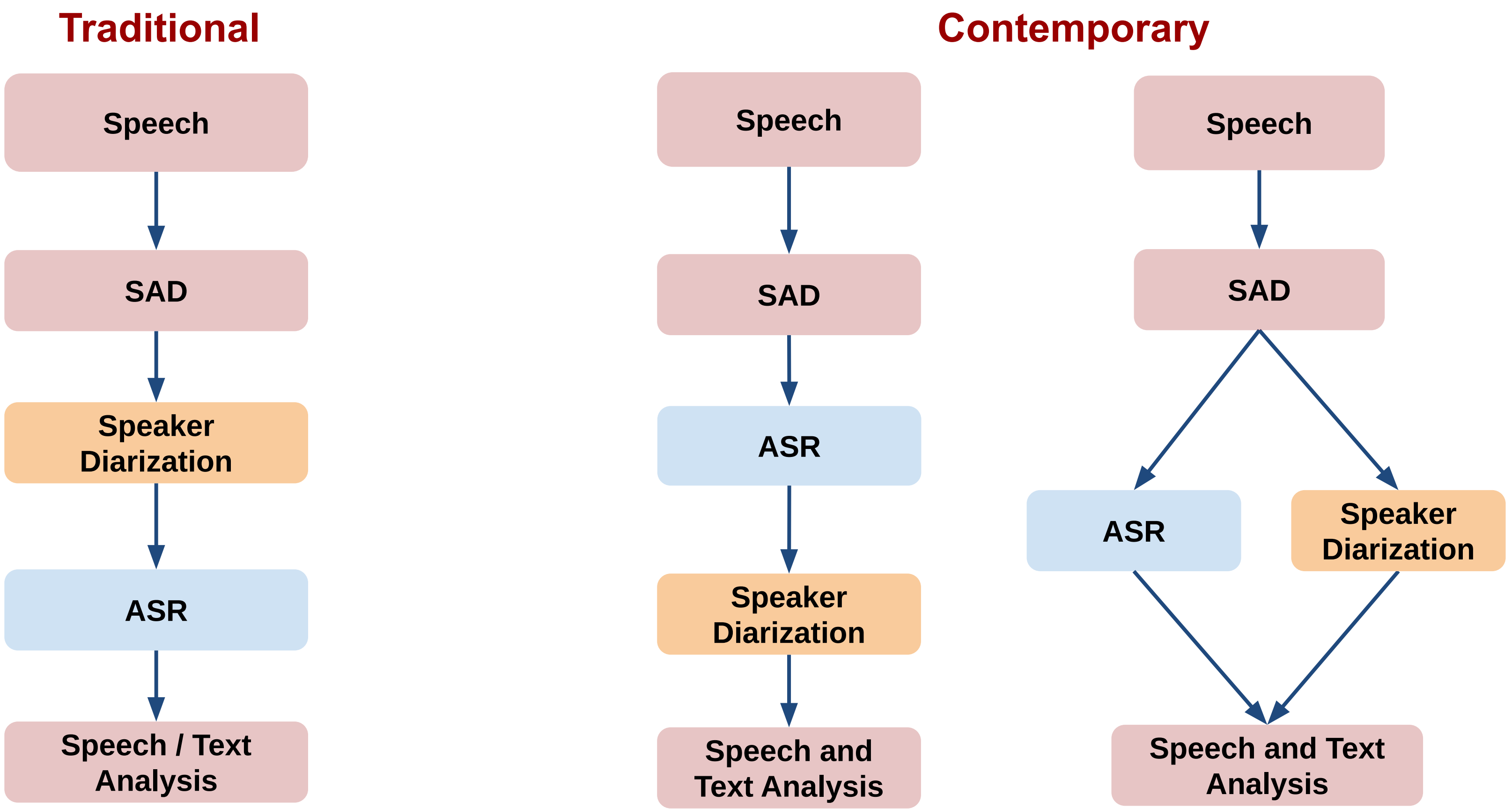
Speaker Diarization and Automatic Speech Recognition

Mismatch between diarization and word boundary from speech recognition



Diarization before speech recognition

Speaker Diarization and Automatic Speech Recognition



Diarization before speech recognition

Leveraging speech recognition for speaker diarization

Chapter 2

Speaker Diarization and ASR

Part-1

Early Studies about Diarization with ASR

Rich Transcription (RT) Evaluation Series

- Purposes
  - Promotes and gauges advances in automatic speech recognition technologies
  - Creates recognition technologies that will produce transcriptions with meta data

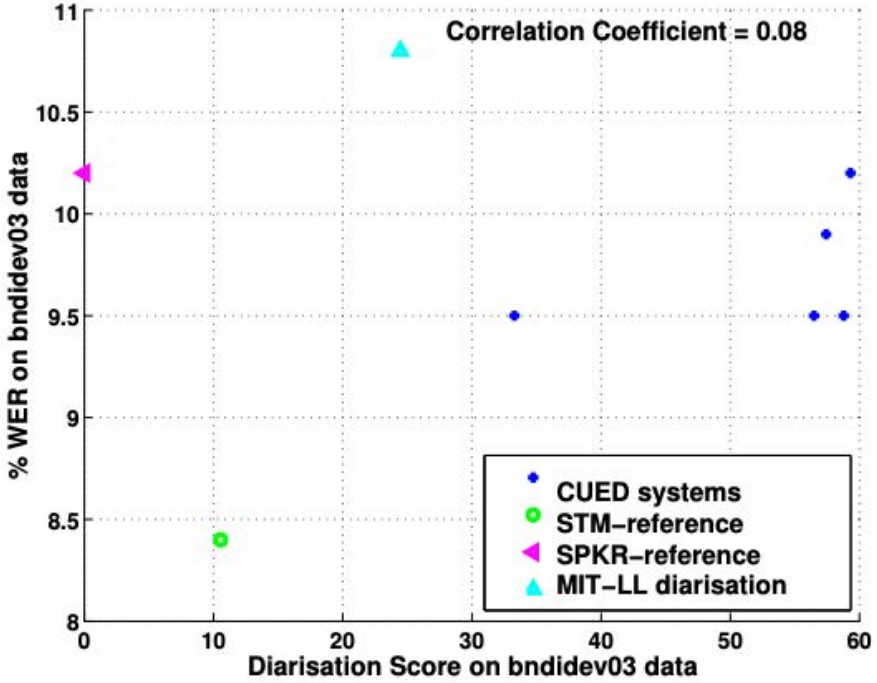
- Main tasks
  - Speech-to-Text Transcription (STT)
    - ASR
  - Metadata Extraction (MDE)
    - Speaker diarization

- Domains / periods
  - CTS, BN and meetings
  - 2002 - 2009

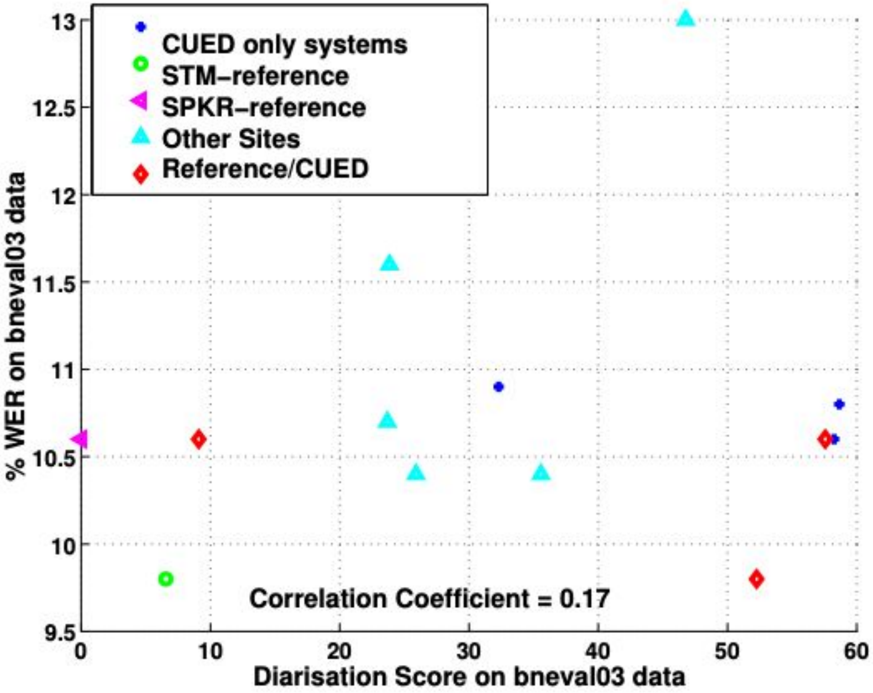
Douglas Reynolds (MIT Lincoln Lab)

Relationship in Error Metrics Between Speaker Diarization and ASR

- Irrelevant!



Correlation coefficient of 0.08.



Correlation coefficient of 0.17.

S. Tranter, et al., “An investigation into the interactions between speaker diarization systems and automatic speech transcription." *CUED/F-INFENG/TR-464*, 2003.

Relationship in Error Metrics Between Speaker Diarization and ASR

- Low diarization error rate (DER) doesn’t guarantee low word error rate (WER).
- Too fine grained boundaries from speaker diarization systems would hurt ASR accuracy.

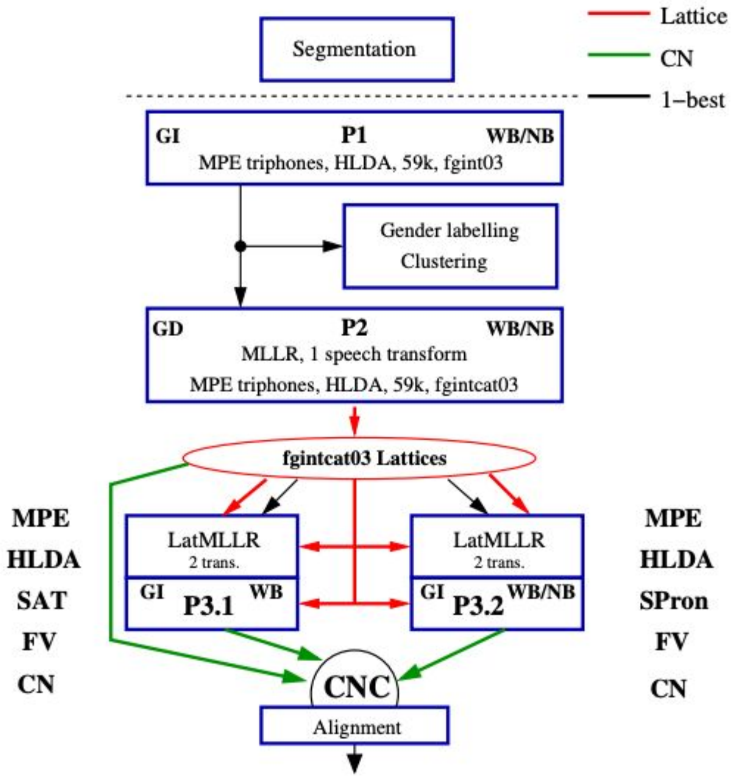
Douglas Reynolds (MIT Lincoln Lab)

Speaker Diarization and Automatic Speech Recognition

Can We Use ASR Outputs to Speaker Diarization for Better WER?

- Diarization outputs vs ASR outputs
  - Segmentation
  - Clustering
  - Recognition

- Baseline ASR system structure for BN
  - Segmentation
  - Speaker clustering
  - Speaker adaptation
  - System combination



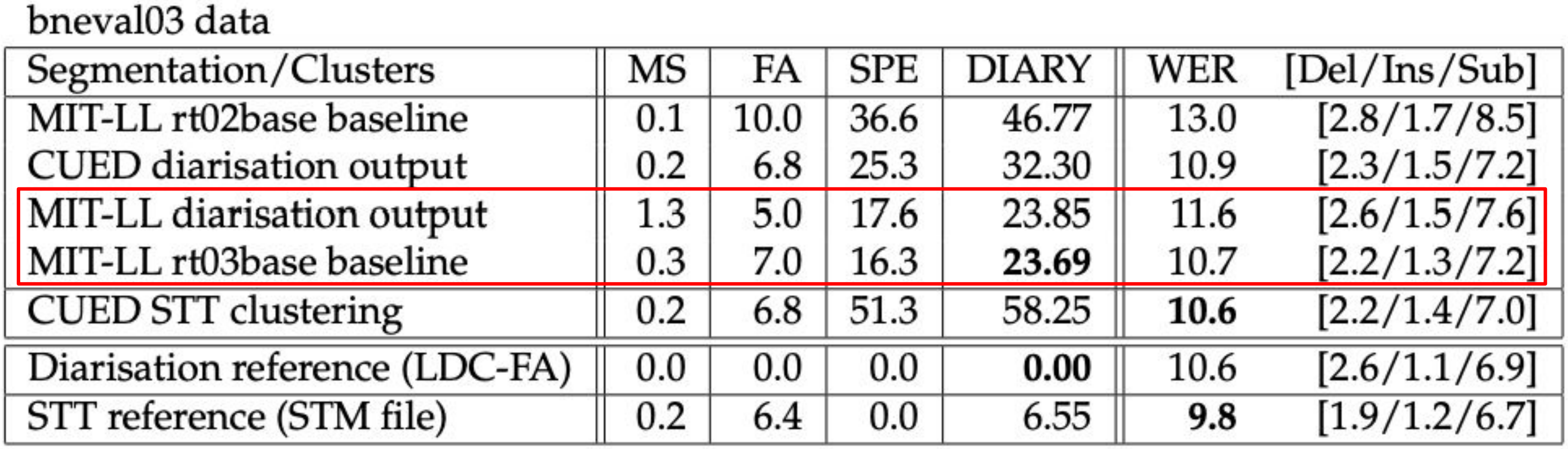
General BN ASR system structure.

S. Tranter, et al., “An investigation into the interactions between speaker diarization systems and automatic speech transcription." *CUED/F-INFENG/TR-464*, 2003.

Speaker Diarization and Automatic Speech Recognition

Can We Use ASR Outputs to Speaker Diarization?

- Missed speech might be better in diarization, but would hurt ASR causing more deletion and substitution errors.



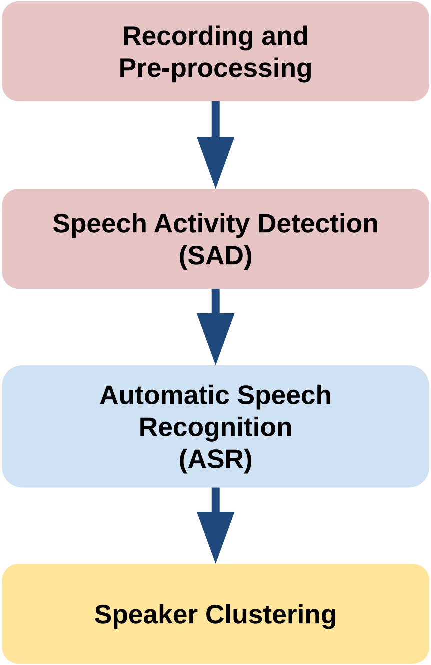
bneval03 data

Segmentation/Clusters	MS	FA	SPE	DIARY	WER	[Del/Ins/Sub]
MIT-LL rt02base baseline	0.1	10.0	36.6	46.77	13.0	[2.8/1.7/8.5]
CUED diarisation output	0.2	6.8	25.3	32.30	10.9	[2.3/1.5/7.2]
MIT-LL diarisation output	1.3	5.0	17.6	23.85	11.6	[2.6/1.5/7.6]
MIT-LL rt03base baseline	0.3	7.0	16.3	**23.69**	10.7	[2.2/1.3/7.2]
CUED STT clustering	0.2	6.8	51.3	58.25	**10.6**	[2.2/1.4/7.0]
Diarisation reference (LDC-FA)	0.0	0.0	0.0	**0.00**	10.6	[2.6/1.1/6.9]
STT reference (STM file)	0.2	6.4	0.0	6.55	**9.8**	[1.9/1.2/6.7]

Effect on using different segmentation / speaker labels for ASR.

S. Tranter, et al., “An investigation into the interactions between speaker diarization systems and automatic speech transcription." *CUED/F-INFENG/TR-464*, 2003.

Refine SAD by Using Word Alignments from ASR


Recording and Pre-processing
Speech Activity Detection (SAD)
Automatic Speech Recognition (ASR)
Speaker Clustering

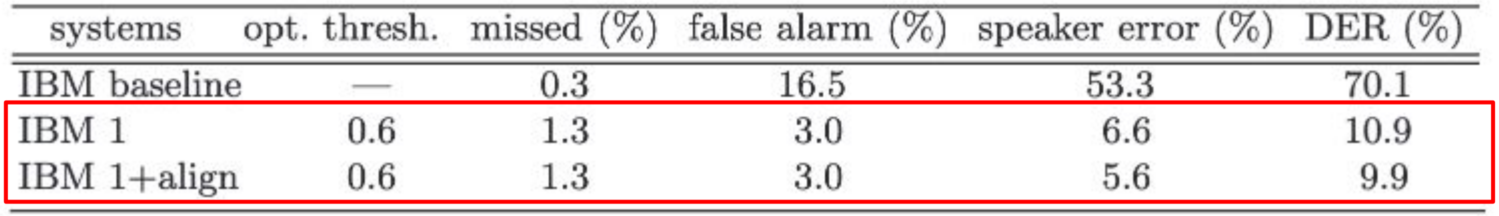
- Missed speech in SAD not recoverable
  - SAD tuned to allow false alarms
  - Segments likely to contain non-speech frames
  - Clustering quality thus degraded
- Incorporates word alignments
  - Uses decoded outputs from a speaker-independent AM to refine SAD results

J. Huang, et al. "The IBM RT07 evaluation systems for speaker diarization on lecture meetings." *Proc. CLEAR / RT*, 2007.

Speaker Diarization and Automatic Speech Recognition

Refine SAD by Using Word Alignments from ASR

- In clustering, frames that correspond to silence, background noise and vocal noise according to ASR word alignments are ignored.



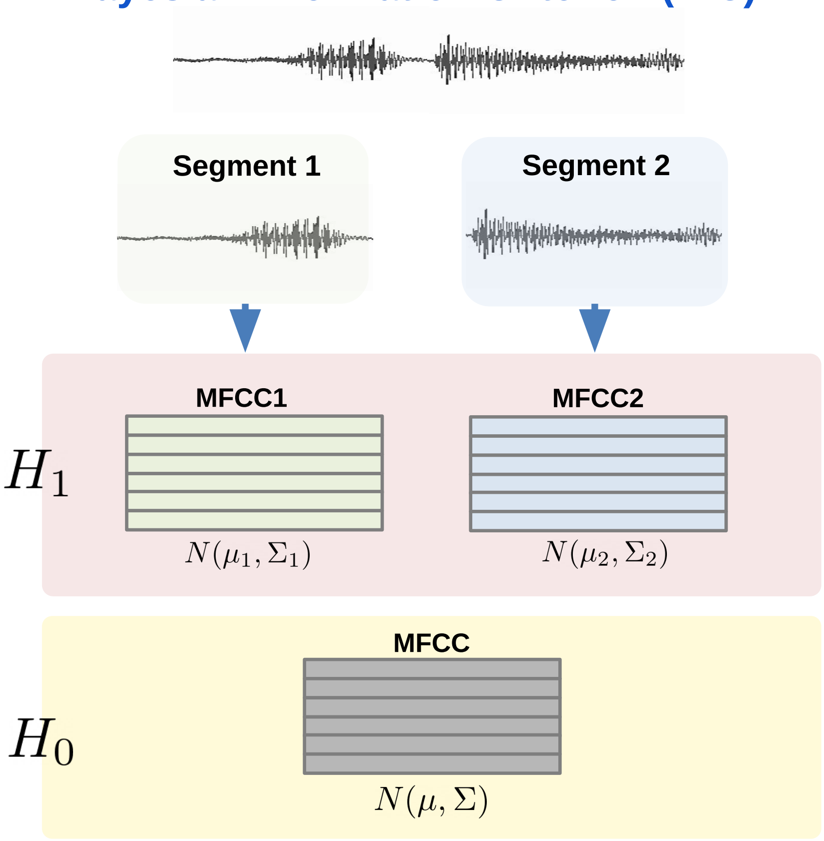
systems	opt. thresh.	missed (%)	false alarm (%)	speaker error (%)	DER (%)
IBM baseline	—	0.3	16.5	53.3	70.1
IBM 1	0.6	1.3	3.0	6.6	10.9
IBM 1+align	0.6	1.3	3.0	5.6	9.9

Diarization error rate break-down.

J. Huang, et al. "The IBM RT07 evaluation systems for speaker diarization on lecture meetings." *Proc. CLEAR / RT*, 2007.

Better Speaker Change Detection by Using Word Alignments from ASR

Speaker Change Detection using Bayesian Information Criterion (BIC)



- Assume a Gaussian process

$$\boldsymbol{x}_i \sim N(\mu_i, \Sigma_i)$$

- Hypothesis testing

$$H_0 : x_1 \cdots x_N \sim N(\mu, \Sigma)$$

$$H_1 : x_1 \cdots x_i \sim N(\mu_1, \Sigma_1)$$

$$x_{i+1} \cdots x_N \sim N(\mu_2, \Sigma_2)$$

- Generalized log likelihood ratio statistic:

$$R = \log\left(\frac{|\Sigma|^N}{|\Sigma_1|^{N_1}|\Sigma_2|^{N_2}}\right)$$

$$= N \log|\Sigma| - N_1 \log|\Sigma_1| - N_2 \log|\Sigma_2|$$

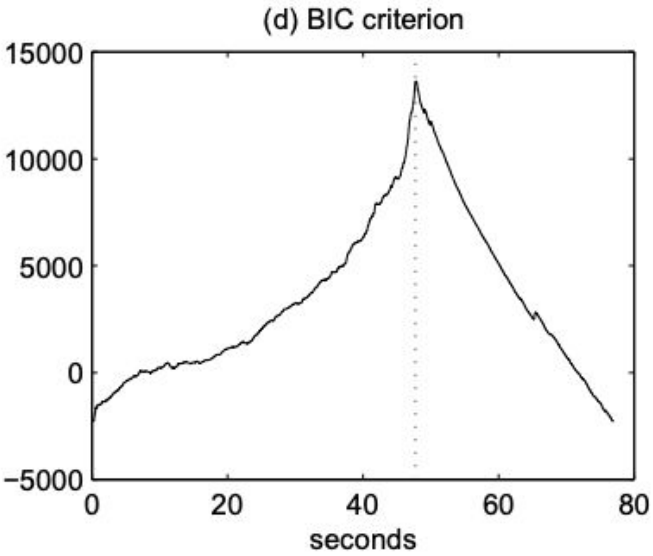
- BIC value

$$BIC = R - \lambda P$$

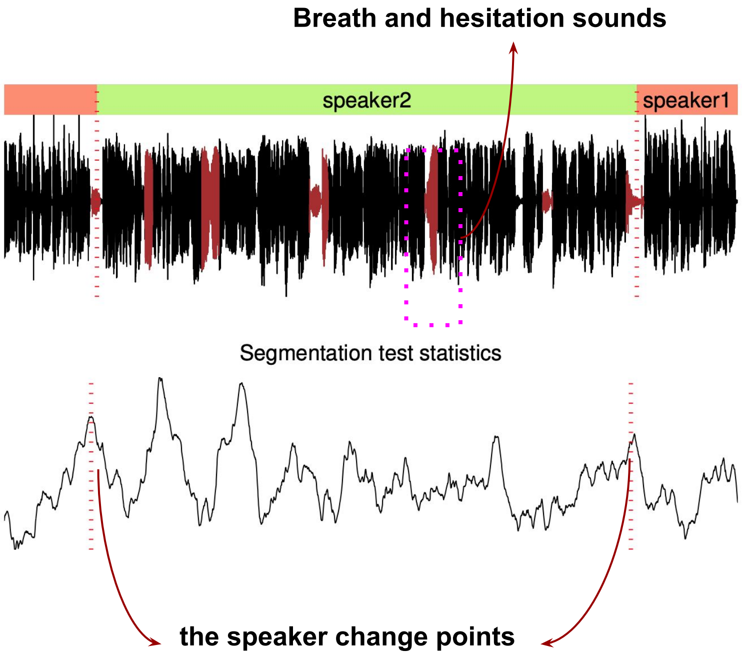
P : model complexity compensation factor

Better Speaker Change Detection by Using Word Alignments from ASR

- Speaker change detection with uniform windowing and BIC
  - Only based on threshold for BIC
  - Not tuned for ASR
  - Very often truncating words



Speaker change detection w/ BIC


Breath and hesitation sounds
speaker2
speaker1
Segmentation test statistics
the speaker change points

Problem of misplaced change points that would cause word truncation

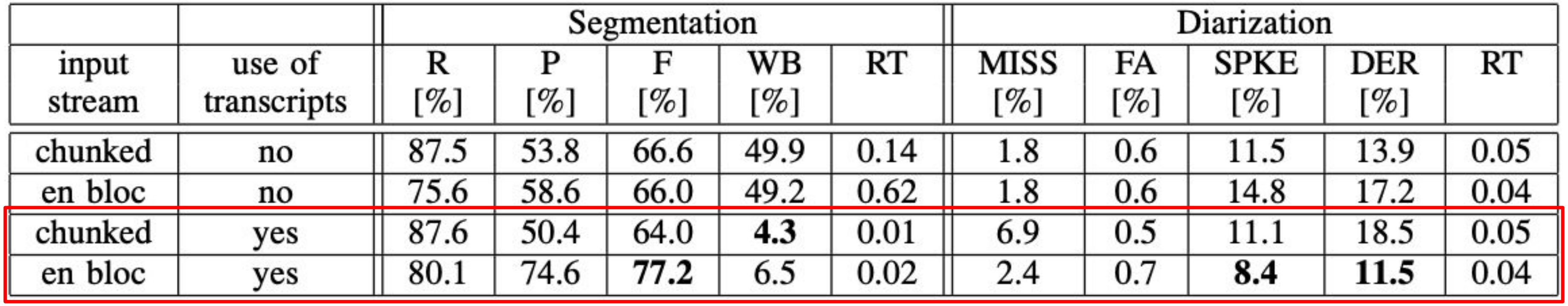
J. Silovsky, et al. "Incorporation of the ASR output in speaker segmentation and clustering within the task of speaker diarization of broadcast streams." *Proc. MMSP*, 2012.

Better Speaker Change Detection by Using Word Alignments from ASR

- Word-breakage (WB)
  - Ratio of change-points that are detected inside intervals corresponding to words (i.e., word truncation)

$$WB = \frac{H_b + I_b}{H + I}$$

H: Number of coupled detections
I: Number of inserted detections
H_b: Number of coupled detections that cause word-breakages
I_b: Number of inserted detections that cause word-breakages



		Segmentation					Diarization				
input stream	use of transcripts	R [%]	P [%]	F [%]	WB [%]	RT	MISS [%]	FA [%]	SPKE [%]	DER [%]	RT
chunked	no	87.5	53.8	66.6	49.9	0.14	1.8	0.6	11.5	13.9	0.05
en bloc	no	75.6	58.6	66.0	49.2	0.62	1.8	0.6	14.8	17.2	0.04
chunked	yes	87.6	50.4	64.0	**4.3**	0.01	6.9	0.5	11.1	18.5	0.05
en bloc	yes	80.1	74.6	**77.2**	6.5	0.02	2.4	0.7	**8.4**	**11.5**	0.04

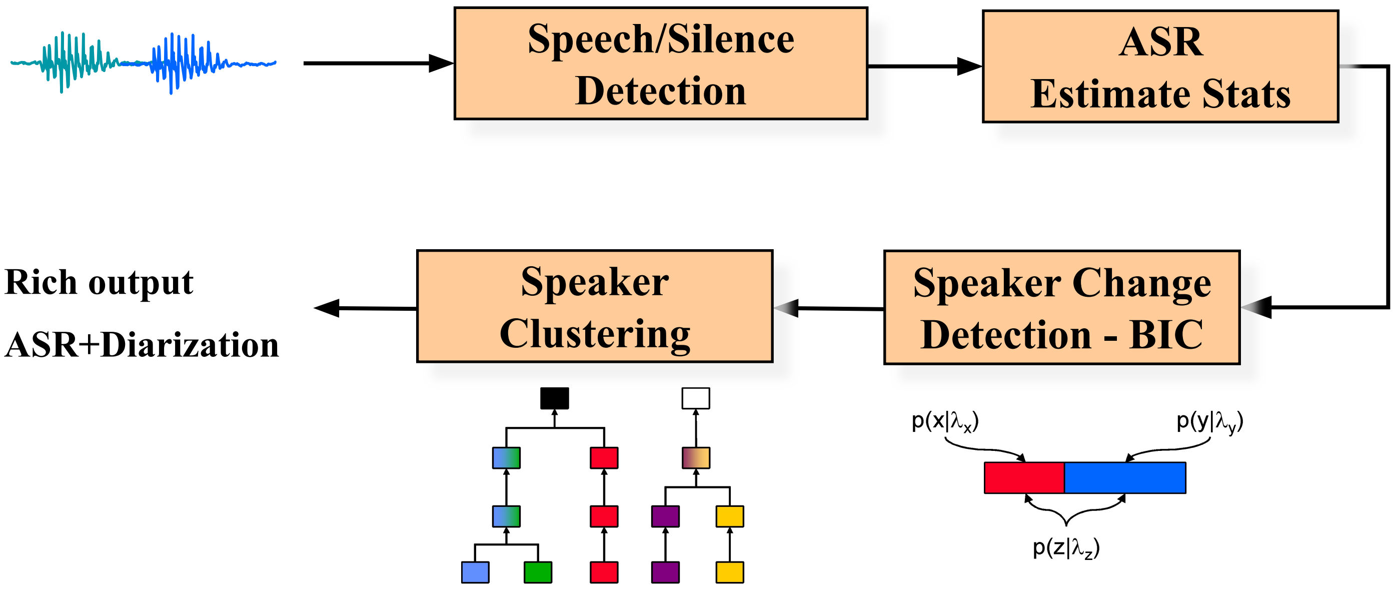
J. Silovsky, et al. "Incorporation of the ASR output in speaker segmentation and clustering within the task of speaker diarization of broadcast streams." *Proc. MMSP*, 2012.

Speaker Diarization and Automatic Speech Recognition

Online speaker diarization using ASR for speaker change point refinement

- Diarization before ASR causing problems
  - Segmentation generating too many false positives or ignoring true speaker turns
  - Tuning possible, but still hard to generalize

- ASR, then diarization!



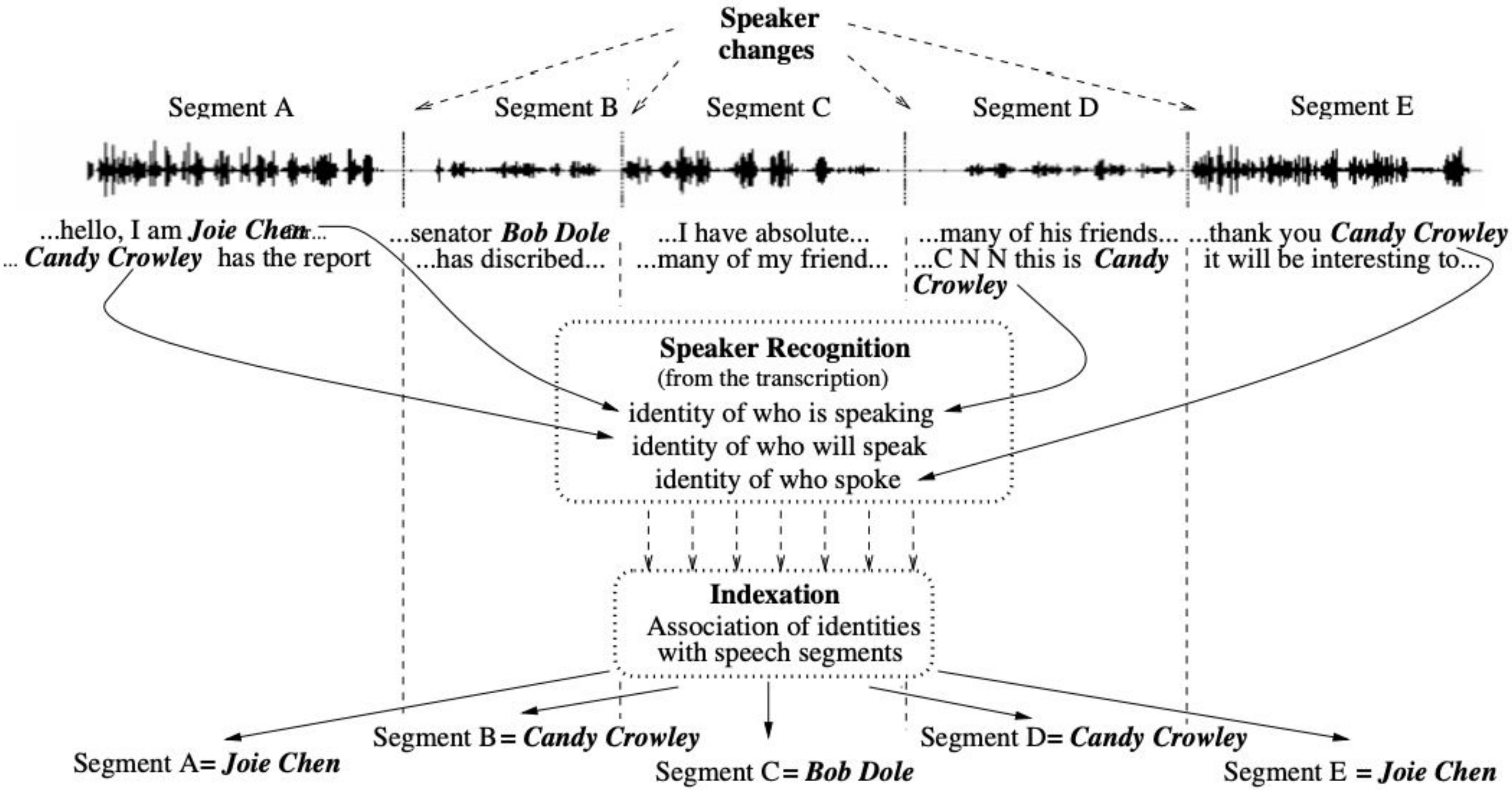
D. Dimitriadis and P. Fousek, "Developing on-line speaker diarization system." *Proc. Interspeech*, 2017.

Speaker Diarization and Automatic Speech Recognition

How about using *linguistic patterns* to identify speakers?

Speaker Diarization and Automatic Speech Recognition

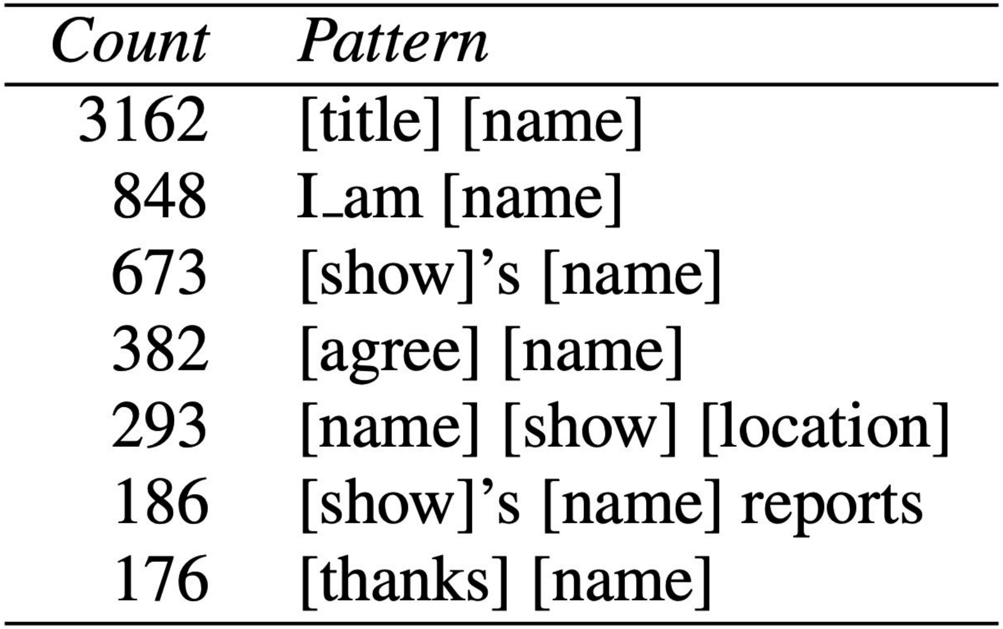
Let’s use *linguistic patterns* to identify speakers!



L. Canseco-Rodriguez, L. Lamel, and J. Gauvain. "Speaker diarization from speech transcripts." *Proc. ICSLP*, 2004.

Speaker Diarization and Automatic Speech Recognition

Linguistic Patterns (from Manual Transcripts)

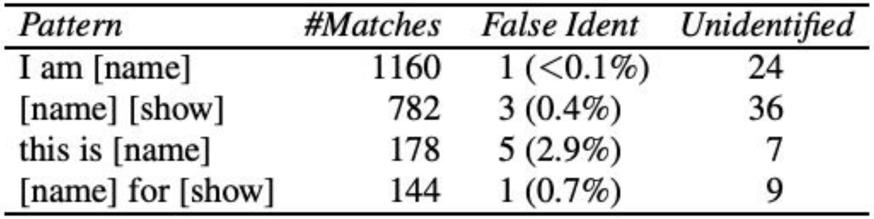


Count	*Pattern*
3162	[title] [name]
848	I_am [name]
673	[show]’s [name]
382	[agree] [name]
293	[name] [show] [location]
186	[show]’s [name] reports
176	[thanks] [name]

Useful patterns to extract speaker identities.

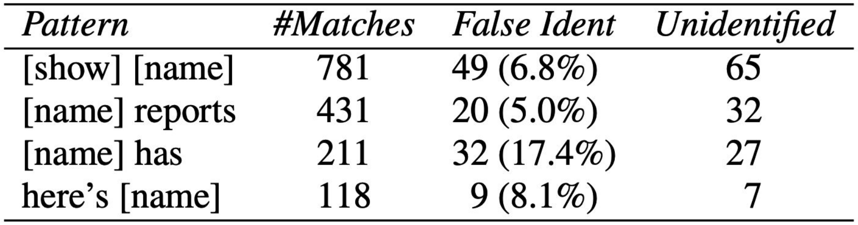
L. Canseco-Rodriguez, L. Lamel, and J. Gauvain. "Speaker diarization from speech transcripts." *Proc. ICSLP*, 2004.

Linguistic Patterns (from Manual Transcripts)



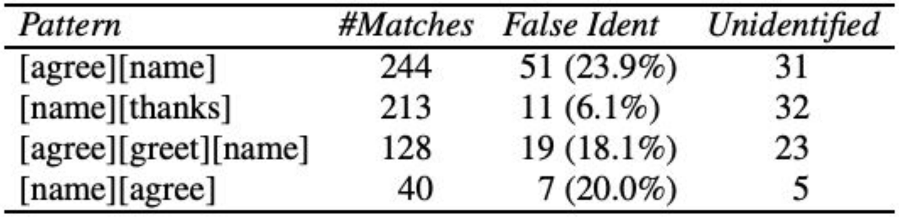
Pattern	*#Matches*	*False Ident*	*Unidentified*
I am [name]	1160	1 (<0.1%)	24
[name] [show]	782	3 (0.4%)	36
this is [name]	178	5 (2.9%)	7
[name] for [show]	144	1 (0.7%)	9

Validation of self-speaker patterns.



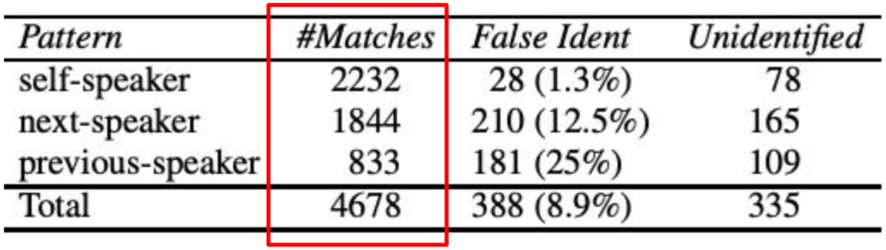
Pattern	*#Matches*	*False Ident*	*Unidentified*
[show] [name]	781	49 (6.8%)	65
[name] reports	431	20 (5.0%)	32
[name] has	211	32 (17.4%)	27
here's [name]	118	9 (8.1%)	7

Validation of next-speaker patterns.



Pattern	*#Matches*	*False Ident*	*Unidentified*
[agree][name]	244	51 (23.9%)	31
[name][thanks]	213	11 (6.1%)	32
[agree][greet][name]	128	19 (18.1%)	23
[name][agree]	40	7 (20.0%)	5

Validation of previous-speaker patterns.



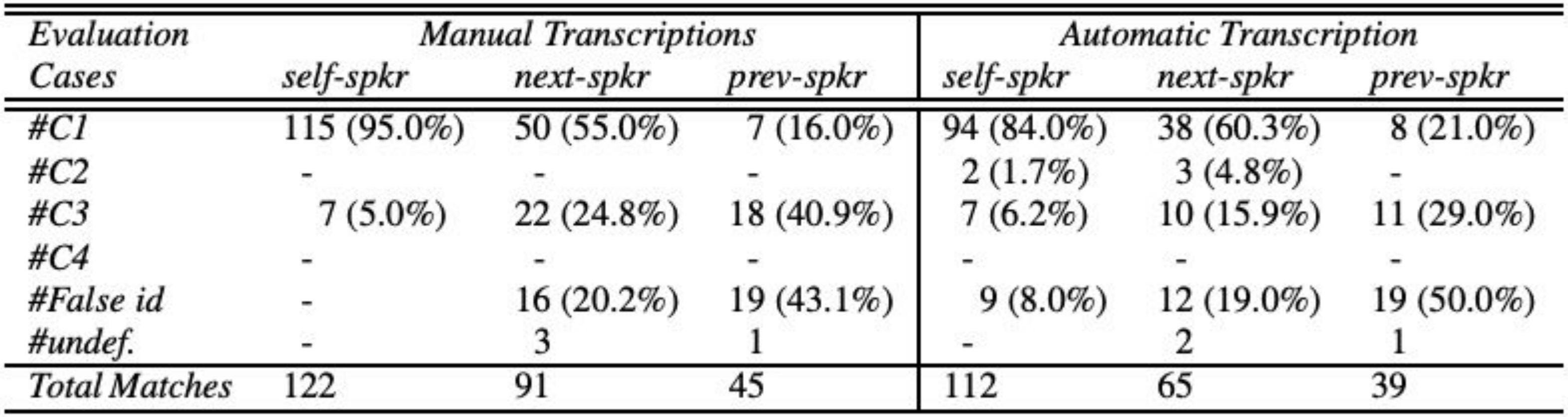
Pattern	*#Matches*	*False Ident*	*Unidentified*
self-speaker	2232	28 (1.3%)	78
next-speaker	1844	210 (12.5%)	165
previous-speaker	833	181 (25%)	109
Total	4678	388 (8.9%)	335

Speaker ID error rates.

L. Canseco-Rodriguez, L. Lamel, and J. Gauvain. "Speaker diarization from speech transcripts." *Proc. ICSLP*, 2004.

Speaker Diarization and Automatic Speech Recognition

Linguistic Patterns (from Automatic Transcripts)



Evaluation	*Manual Transcriptions*			*Automatic Transcription*		
Cases	*self-spkr*	*next-spkr*	*prev-spkr*	*self-spkr*	*next-spkr*	*prev-spkr*
#C1	115 (95.0%)	50 (55.0%)	7 (16.0%)	94 (84.0%)	38 (60.3%)	8 (21.0%)
#C2	-	-	-	2 (1.7%)	3 (4.8%)	-
#C3	7 (5.0%)	22 (24.8%)	18 (40.9%)	7 (6.2%)	10 (15.9%)	11 (29.0%)
#C4	-	-	-	-	-	-
#False id	-	16 (20.2%)	19 (43.1%)	9 (8.0%)	12 (19.0%)	19 (50.0%)
#undef.	-	3	1	-	2	1
Total Matches	122	91	45	112	65	39

Diarization rates on eval data.

#C1: Identity associated with pure speaker turn, matching reference identity

#C2: Identity associated with impure speaker turn, matching reference identity

#C3: Identity associated with pure speaker turn, partially matching reference identity

#C4: Identity associated with impure speaker turn, partially matching reference identity

#undef.: Identity matching unidentified speaker in reference

#False id: None of above, erroneous identity association

L. Canseco, L. Lamel, and J. Gauvain. "A comparative study using manual and automatic transcripts for diarization." *Proc. ASRU*, 2005.

Still, Not Fully Benefiting from Linguistic Information

- Language model style approach helpful for diarization
- Current diarization systems, lacking such modeling to understand what people say and how they take turns

Andreas Stolcke (Amazon)

Chapter 2

Speaker Diarization and ASR

Part-2

Lexical Information Used in Speaker Diarization

Lexical Information Used in Speaker Diarization

Speaker diarization and lexical feature

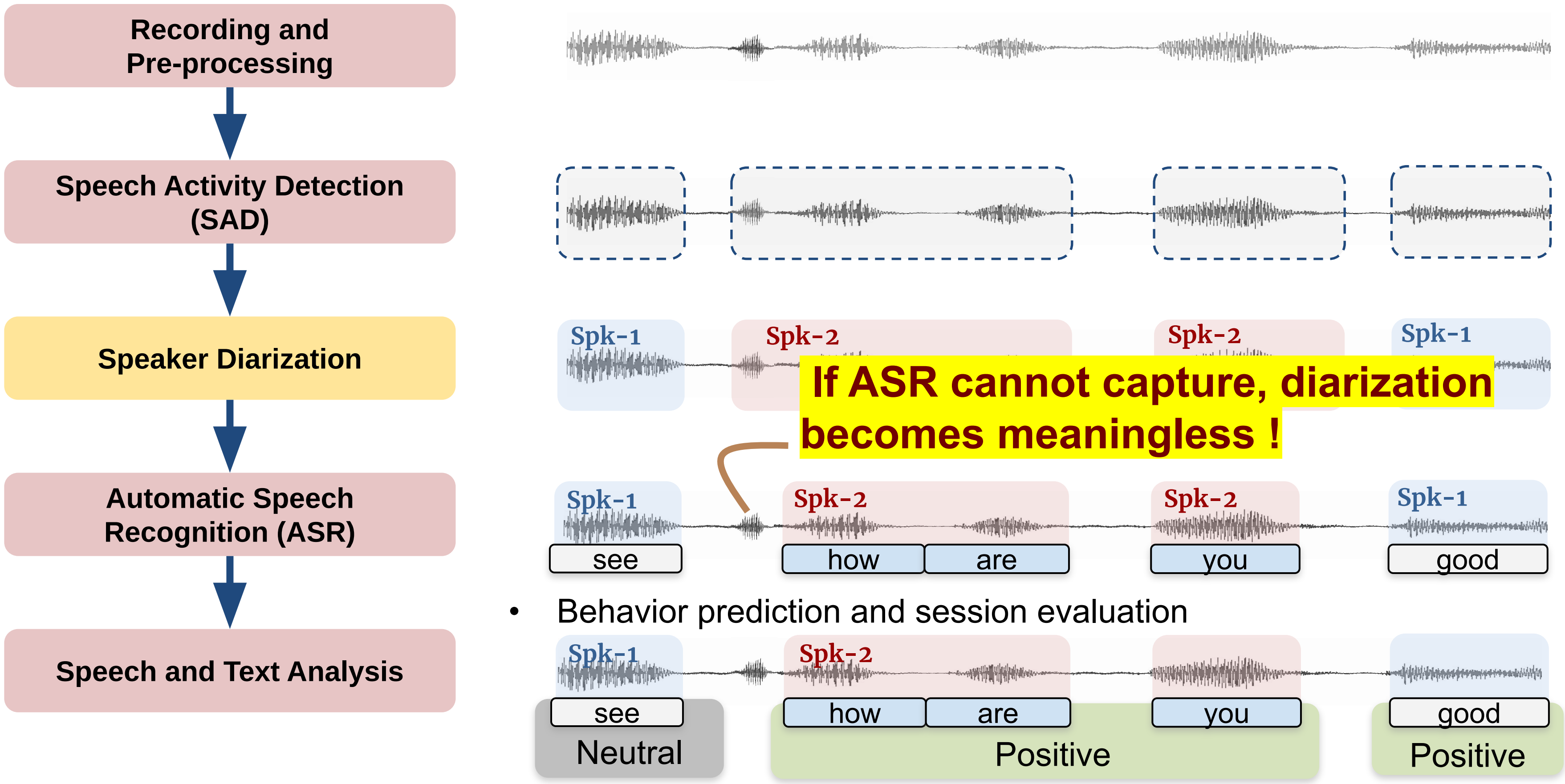
Lexical feature for speaker diarization

- Lexical feature often contains topic information or speaker specific pattern
- Lexical information can compensate the sparse acoustic information from a specific speaker.
- Lexical approach can only be useful when ASR and segmentation outputs are reliable.

Katrin Kirchhoff (Amazon)

Lexical Information Used in Speaker Diarization

Motivation: Speech Processing Pipeline

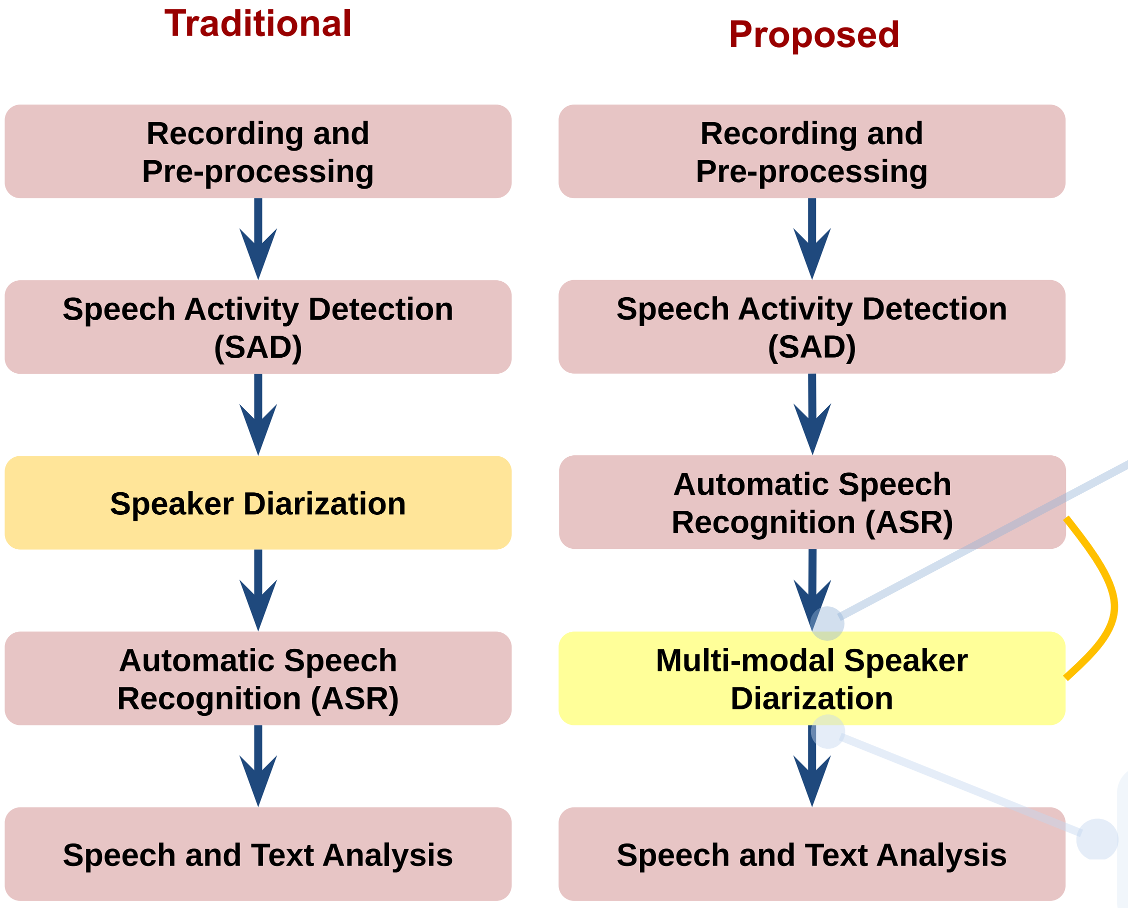


If ASR cannot capture, diarization becomes meaningless !

- Behavior prediction and session evaluation

Lexical Information Used in Speaker Diarization

Motivation: Speech Processing Pipeline


Traditional
Recording and Pre-processing
Speech Activity Detection (SAD)
Speaker Diarization
Automatic Speech Recognition (ASR)
Speech and Text Analysis
Proposed
Recording and Pre-processing
Speech Activity Detection (SAD)
Automatic Speech Recognition (ASR)
Multi-modal Speaker Diarization
Speech and Text Analysis

Why after ASR?

- Outside ASR output, diarization becomes pointless
- Lexical information (words) can help figuring

After ASR

how is your day going quite busy you must feel stressed out

Lexical Info (words) + Time Stamps

After Diarization

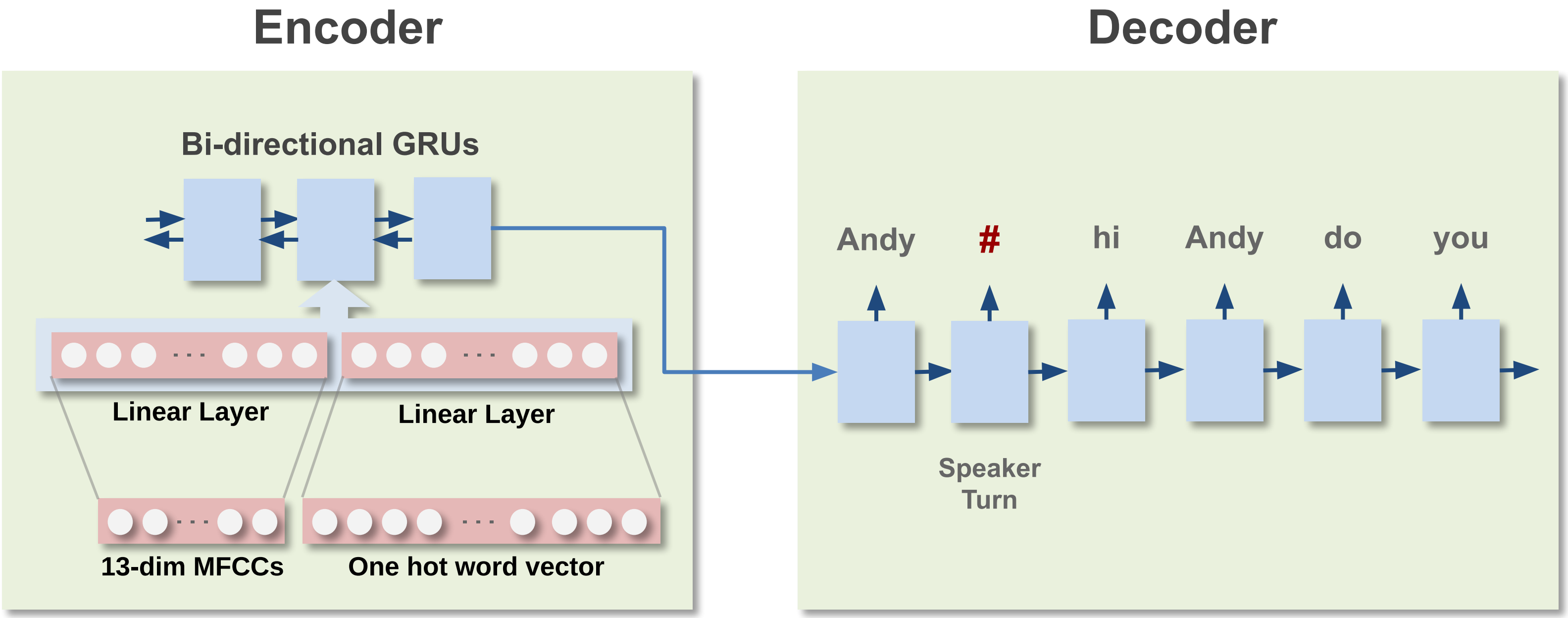
Speaker A: how is your day going
Speaker B: quite busy
Speaker A: you must feel stressed out

Park, Tae Jin, and Panayiotis Georgiou. "Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks." arXiv preprint arXiv:1805.10731 (2018).

Lexical Information Used in Speaker Diarization

Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks (Park et. al.)

Sequence to sequence: Encoder and Decoder

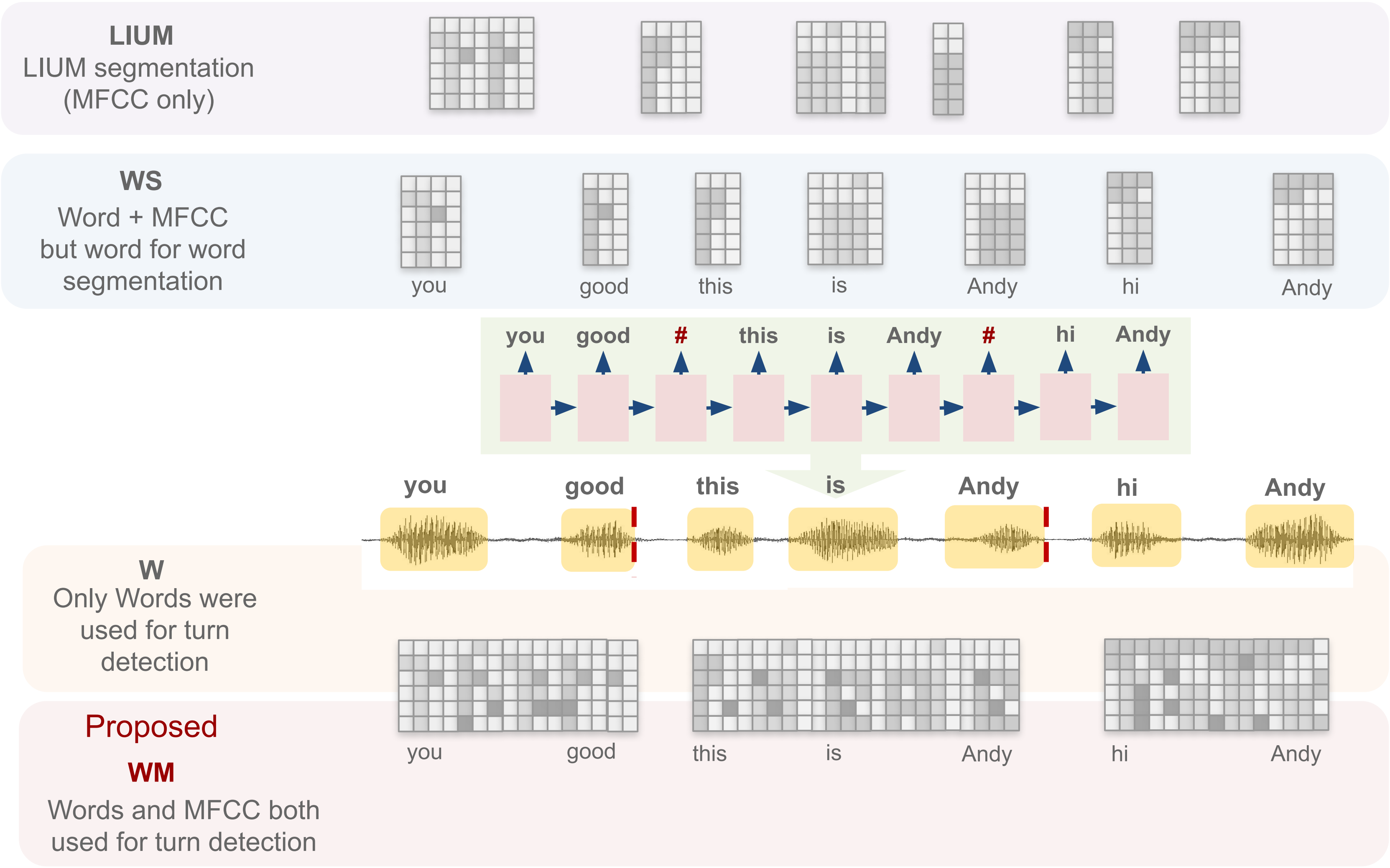


- Encoder processes both acoustic and lexical input and hand over to the decoder.
- Decoder outputs turn tokens (#) with the original input sentence.

Park, Tae Jin, and Panayiotis Georgiou. "Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks." arXiv preprint arXiv:1805.10731 (2018).

Lexical Information Used in Speaker Diarization

Majority Vote and Turn Decision



LIUM
LIUM segmentation
(MFCC only)

WS
Word + MFCC
but word for word
segmentation

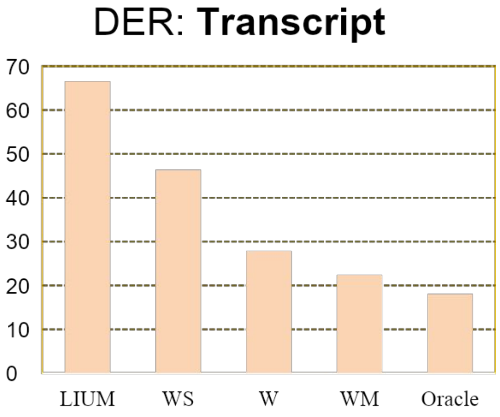
W
Only Words were
used for turn
detection

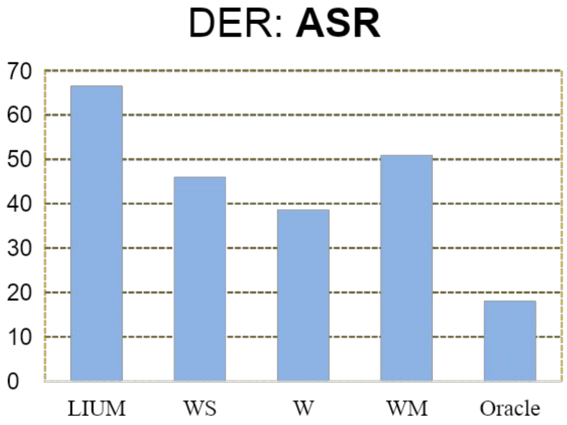
Proposed
WM
Words and MFCC both
used for turn detection

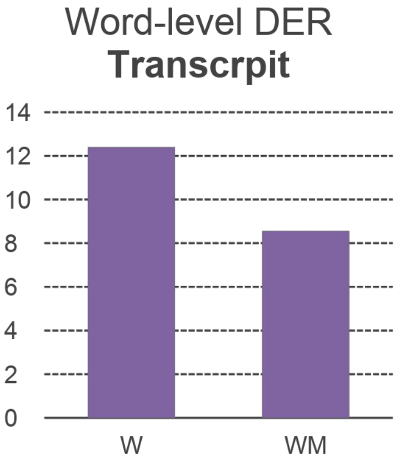
Park, Tae Jin, and Panayiotis Georgiou. "Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks." arXiv preprint arXiv:1805.10731 (2018).

Lexical Information Used in Speaker Diarization

Experimental Results: The effect of ASR performance on segmentation







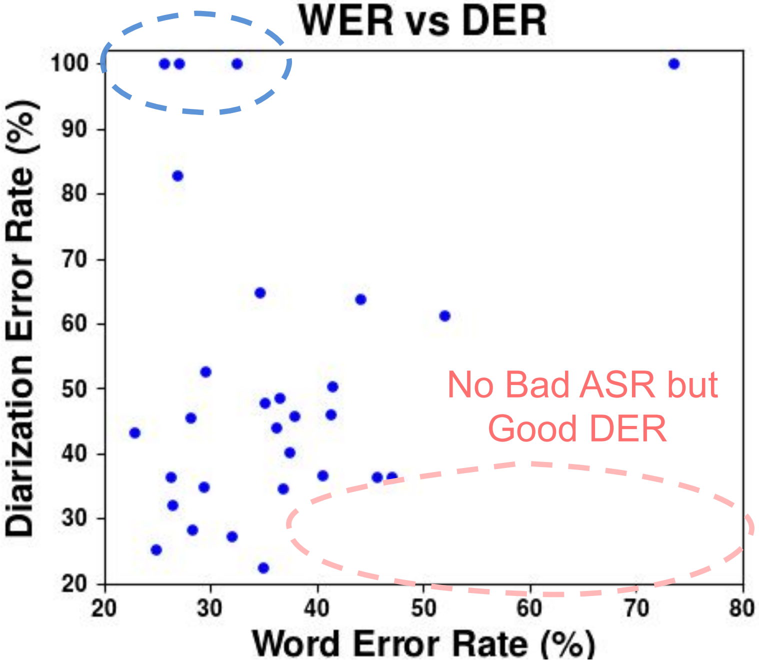
- With transcript, WM model showed the best performance
- With ASR, WM model did not perform well while W model still out-performs others
- WDER = $\frac{\#\ of\ Correctly\ Diarized\ Words}{Total\ \#\ of\ Words}$
- WDER reflects the actual diarization result we see

Park, Tae Jin, and Panayiotis Georgiou. "Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks." arXiv preprint arXiv:1805.10731 (2018).

Lexical Information Used in Speaker Diarization

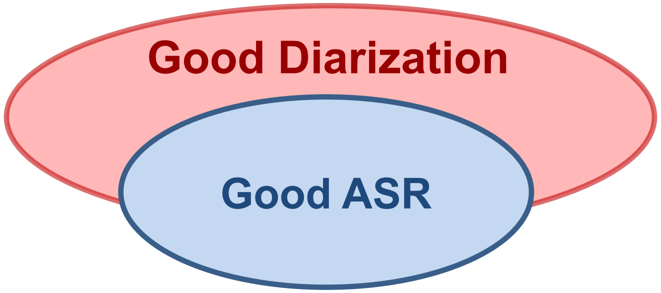
Experimental Results: The effect of ASR performance on segmentation

Some Good ASR
but Bad DER

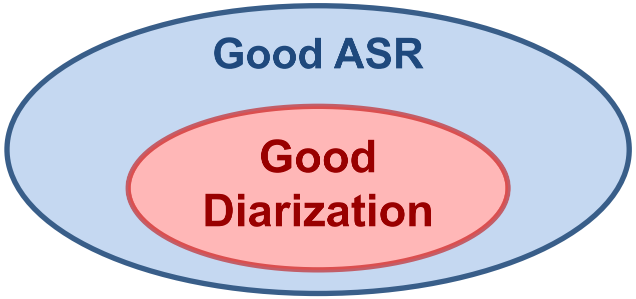


No Bad ASR but
Good DER

In general, good Diarization does NOT necessarily lead to good ASR results.



If we use ASR(Lexical) result for segmentation and diarization, then:

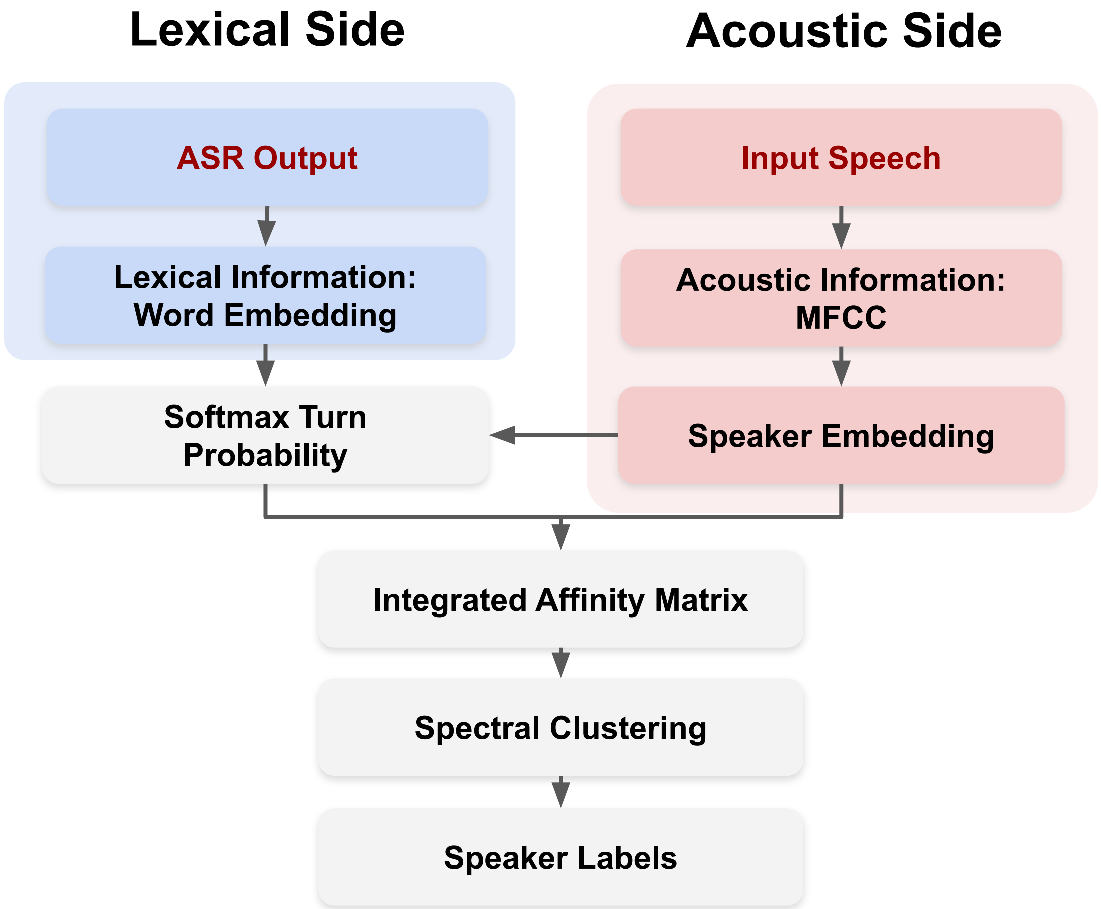


- Even if ASR result was good, some sessions are very hard to get good performances
- If ASR results are bad, DER results are usually poor if we use ASR result for diarization.

Park, Tae Jin, and Panayiotis Georgiou. "Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks." arXiv preprint arXiv:1805.10731 (2018).

Lexical Information Used in Speaker Diarization

Speaker Diarization with Lexical Information (Park et. al)



Lexical Side

Acoustic Side

ASR Output

Lexical Information: Word Embedding

Input Speech

Acoustic Information: MFCC

Speaker Embedding

Softmax Turn Probability

Integrated Affinity Matrix

Spectral Clustering

Speaker Labels

Lexical Information Used in Speaker Diarization

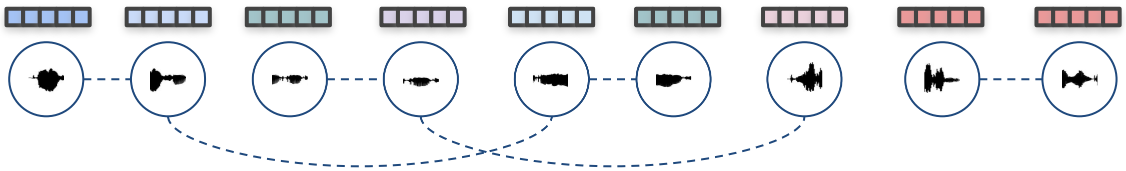
Speaker Diarization with Lexical Information (Park et. al)

	1	2	3	4	5	6	7	8
1	1	1	0	0	0	0	1	1
2	1	1	0	0.5	0	0	1	0.5
3	0	0	1	0	1	0.5	0.5	0
4	0	0.5	0	1	0.5	1	0	0
5	0	0	1	0.5	1	1	0.5	0
6	0	0	0.5	1	1	1	0.5	0.5
7	1	1	0.5	0	0.5	0.5	1	1
8	1	0.5	0	0	0	0.5	1	1

Adjacency matrix $\mathbf{P}_{ud}$
from speaker embeddings

Graph Perspective of Speaker Diarization: Spectral Clustering

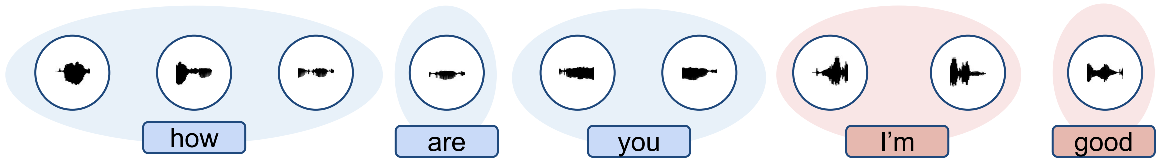
Speaker Embedding Similarity Graph



		good	good	how	are	you	you	still	still
		1	2	3	4	5	6	7	8
good	1	0	0	0	0	0	0	0	0
good	2	0	0	0	0	0	0	0	0
how	3	0	0	1	1	1	1	0	0
are	4	0	0	1	1	1	1	0	0
you	5	0	0	1	1	1	1	0	0
you	6	0	0	1	1	1	1	0	0
still	7	0	0	0	0	0	0	0	0
still	8	0	0	0	0	0	0	0	0

Adjacency matrix $\mathbf{Q}_{c}$
from speaker turn estimations

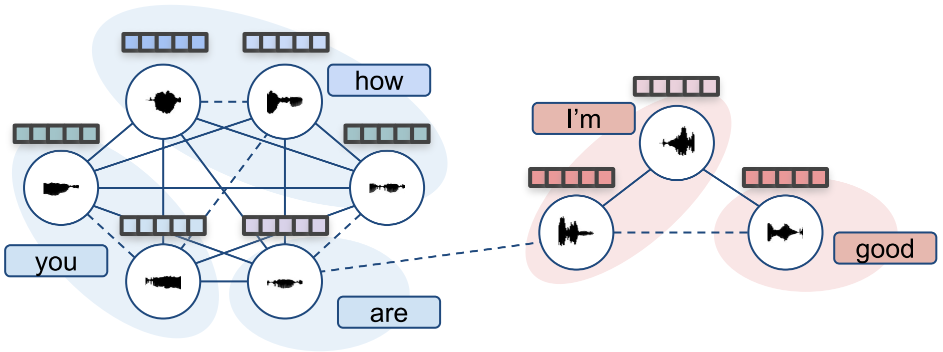
Speech Recognition Output



		good	good	how	are	you	you	doing	doing
		1	2	3	4	5	6	7	8
good	1	1	1	0	0	0	0	1	1
good	2	1	1	0	0.5	0	0	1	0.5
how	3	0	0	1	1	1	1	0.5	0
are	4	0	0.5	1	1	1	1	0	0
you	5	0	0	1	1	1	1	0.5	0
you	6	0	0	1	1	1	1	0.5	0.5
doing	7	1	1	0.5	0	0.5	0.5	1	1
doing	8	1	0.5	0	0	0	0.5	1	1

Integrated Adjacency matrix $\mathbf{A}_{c}$

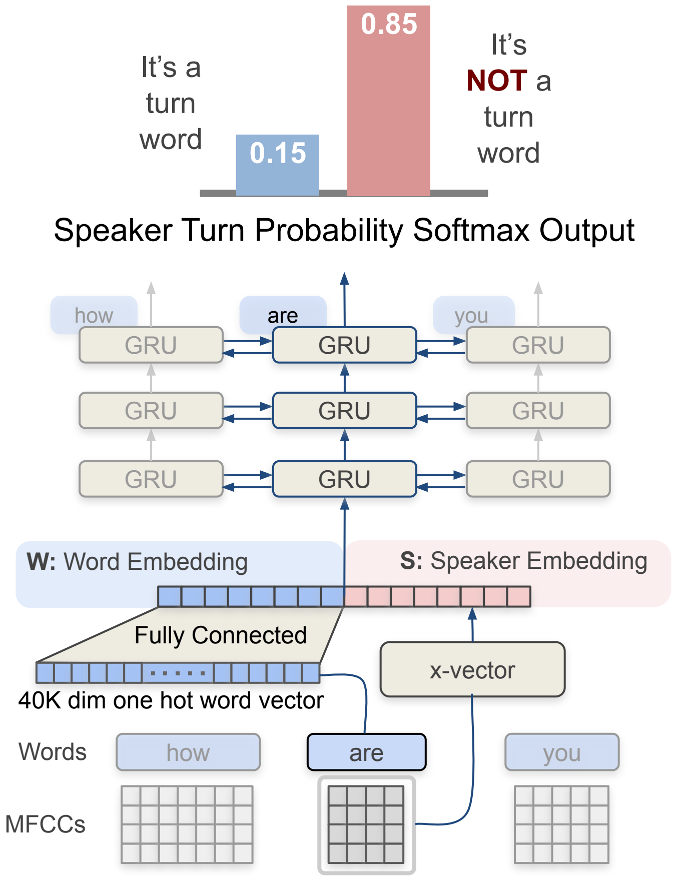
Integrated Similarity Graph



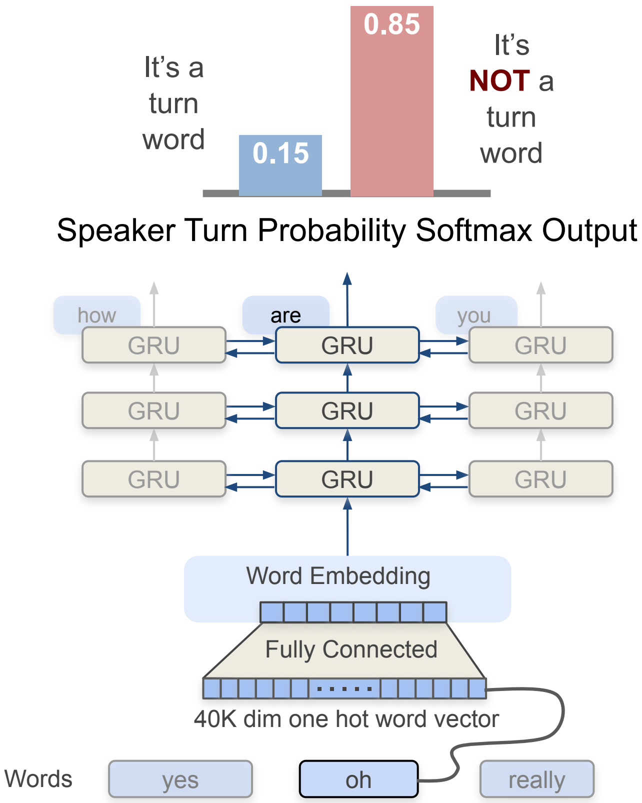
Lexical Information Used in Speaker Diarization

Speaker Diarization with Lexical Information (Park et. al)

Speaker Turn Estimation



Word and Speaker Embedding



Word only

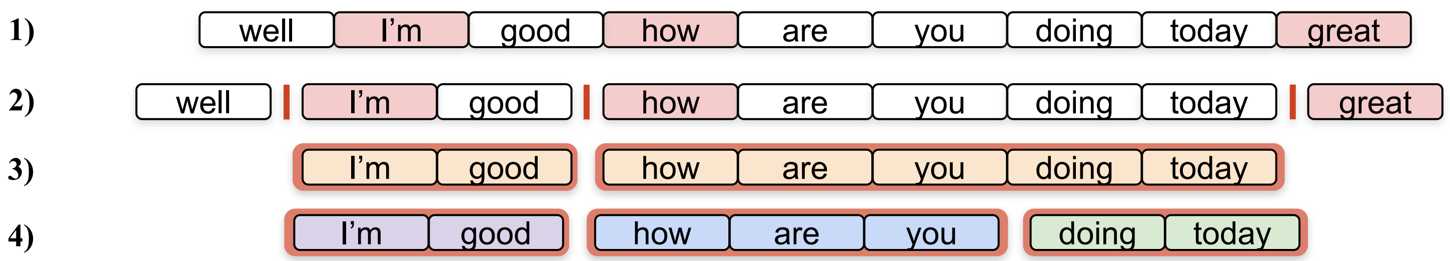
Lexical Information Used in Speaker Diarization

Speaker Diarization with Lexical Information (Park et. al)

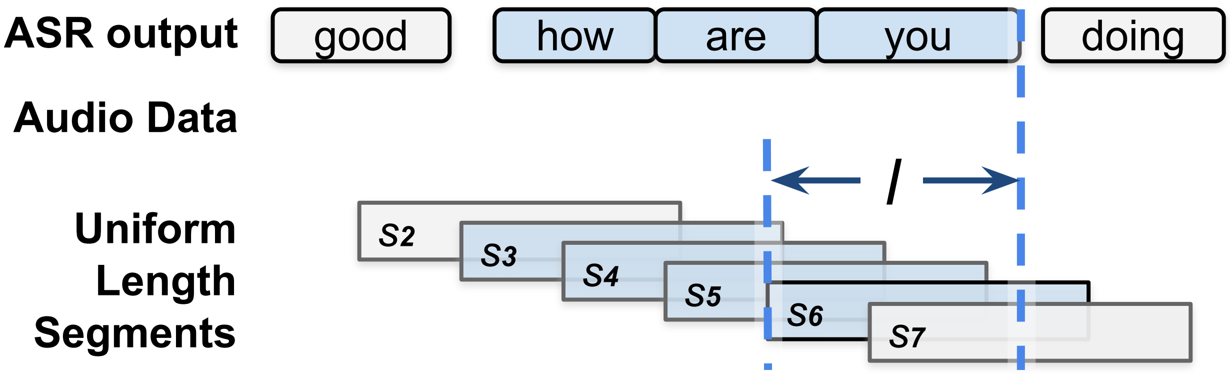
Turn Probability Estimation

Threshold $c = 0.3$, Maximum Utterance Length $\nu = 3$

Turn Probability	0.15	**0.74**	0.06	**0.42**	0.06	0.21	0.03	0.26	**0.34**
Word Sequence	well	I'm	good	how	are	you	doing	today	great

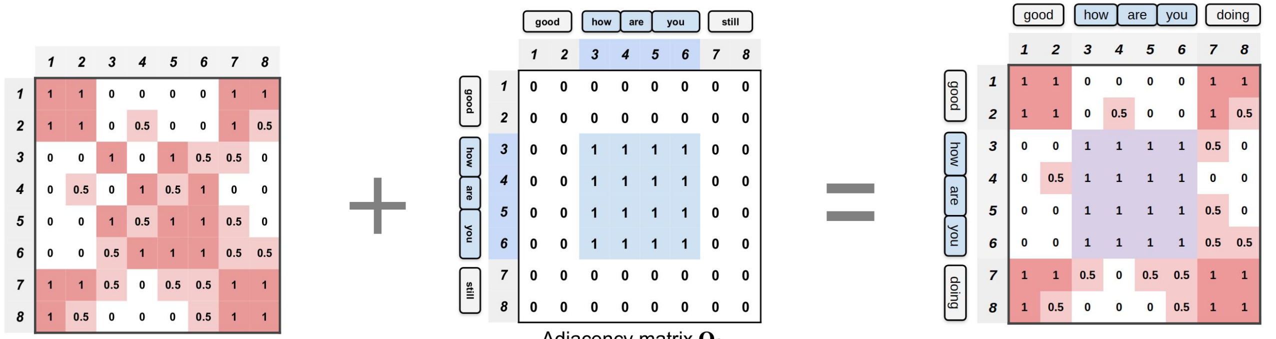


- ☐ Words from ASR
- 🟥 Turn Words
- 🟧 Selected Words
- 🟥 Utterance



Lexical Information Used in Speaker Diarization

Fusion of The Two Affinity Matrices



Acoustic Information

	1	2	3	4	5	6	7	8
1	1	1	0	0	0	0	1	1
2	1	1	0	0.5	0	0	1	0.5
3	0	0	1	0	1	0.5	0.5	0
4	0	0.5	0	1	0.5	1	0	0
5	0	0	1	0.5	1	1	0.5	0
6	0	0	0.5	1	1	1	0.5	0.5
7	1	1	0.5	0	0.5	0.5	1	1
8	1	0.5	0	0	0	0.5	1	1

Adjacency matrix $\mathbf{P}_{\mathbf{ud}}$
from speaker embeddings

+

Lexical Information

		good	good	how	are	you	you	still	still
		1	2	3	4	5	6	7	8
good	1	0	0	0	0	0	0	0	0
good	2	0	0	0	0	0	0	0	0
how	3	0	0	1	1	1	1	0	0
are	4	0	0	1	1	1	1	0	0
you	5	0	0	1	1	1	1	0	0
you	6	0	0	1	1	1	1	0	0
still	7	0	0	0	0	0	0	0	0
still	8	0	0	0	0	0	0	0	0

Adjacency matrix $\mathbf{Q}_{\mathbf{c}}$
from speaker turn estimations

=

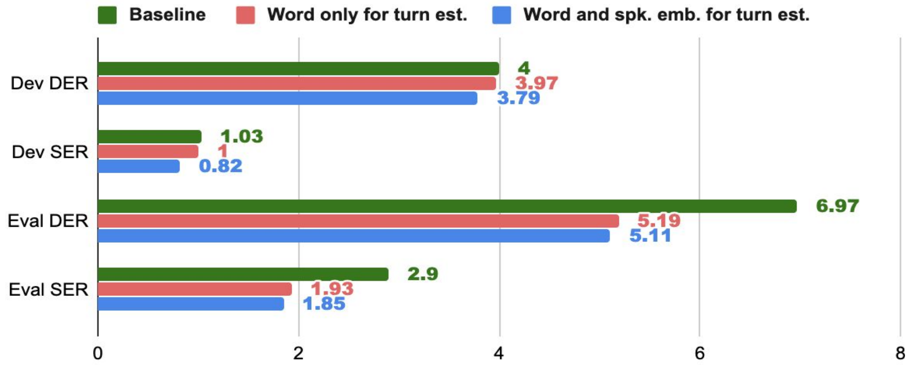
Fused Information

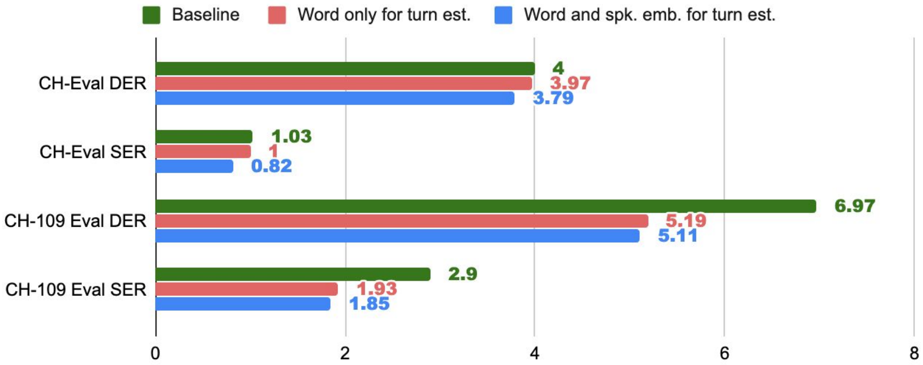
		good	good	how	are	you	you	doing	doing
		1	2	3	4	5	6	7	8
good	1	1	1	0	0	0	0	1	1
good	2	1	1	0	0.5	0	0	1	0.5
how	3	0	0	1	1	1	1	0.5	0
are	4	0	0.5	1	1	1	1	0	0
you	5	0	0	1	1	1	1	0.5	0
you	6	0	0	1	1	1	1	0.5	0.5
doing	7	1	1	0.5	0	0.5	0.5	1	1
doing	8	1	0.5	0	0	0	0.5	1	1

Integrated Adjacency matrix $\mathbf{A}_{\mathbf{c}}$

- Adjacency matrix integration with max operator:

$$\mathbf{A_c} = \max\left(\mathbf{P_{ud}}, \mathbf{Q_c}\right) = \max\left(\frac{1}{2}(\mathbf{P} + \mathbf{P}^{\mathbf{T}}), \mathbf{Q_c}\right)$$





*__SER__: Speaker Error Rate (Confusion) – other than miss or false positive
*__DER__: Diarization Error Rate

Chapter 2

Speaker Diarization and ASR

Part-3

Joint Modeling of Speaker Diarization and ASR

Joint ASR + SD

Why Do We Need Joint Modeling of ASR and SD?

- Joint modeling approach can be one solution to the decoupling of two systems.

- It can take the benefit of utilizing mutual dependency between speaker diarization and ASR.

Naoyuki Kanda (Microsoft)

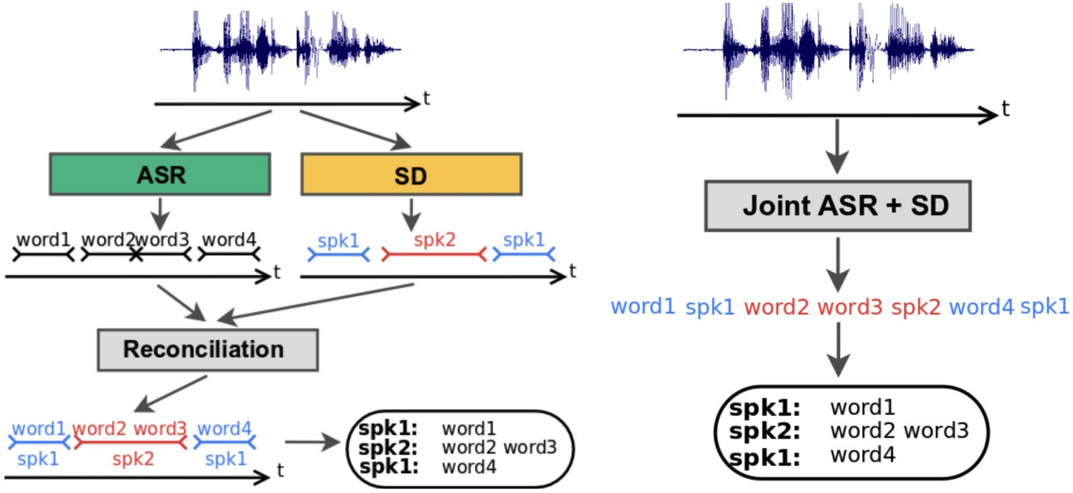
Joint ASR + SD

RNN-T for Sequence Transduction of ASR and SD

Conventional vs joint ASR+SD

- Reconciliation (in labeling and timestamping) between ASR outputs and SD outputs needed in conventional methods

- Joint ASR+SD via sequence transduction, innately dealing with the reconciliation challenges from a sequence labeling perspective



(a) *Baseline*

(b) *Joint ASR+SD*

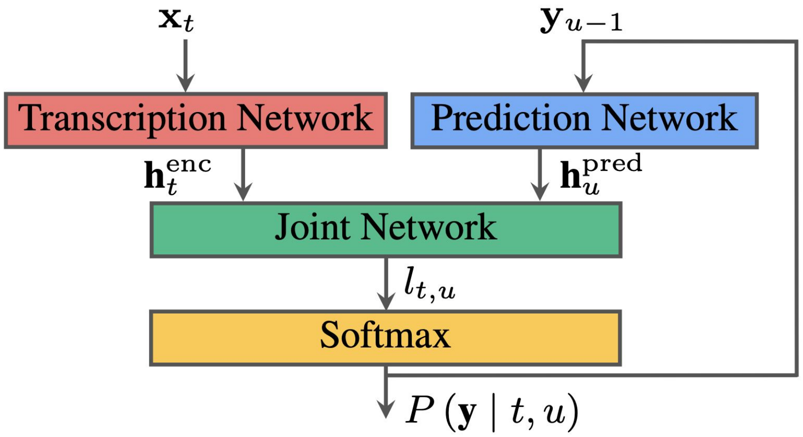
L. Shafey, et al. "Joint speech recognition and speaker diarization via sequence transduction." *Proc. Interspeech*, 2019.

Joint ASR + SD

RNN-T for Sequence Transduction of ASR and SD

hello dr jekyll `<spk:pt>` hello mr hyde what brings you here today `<spk:dr>` I am struggling again with my bipolar disorder `<spk:pt>`

Data example augmented with speaker roles.

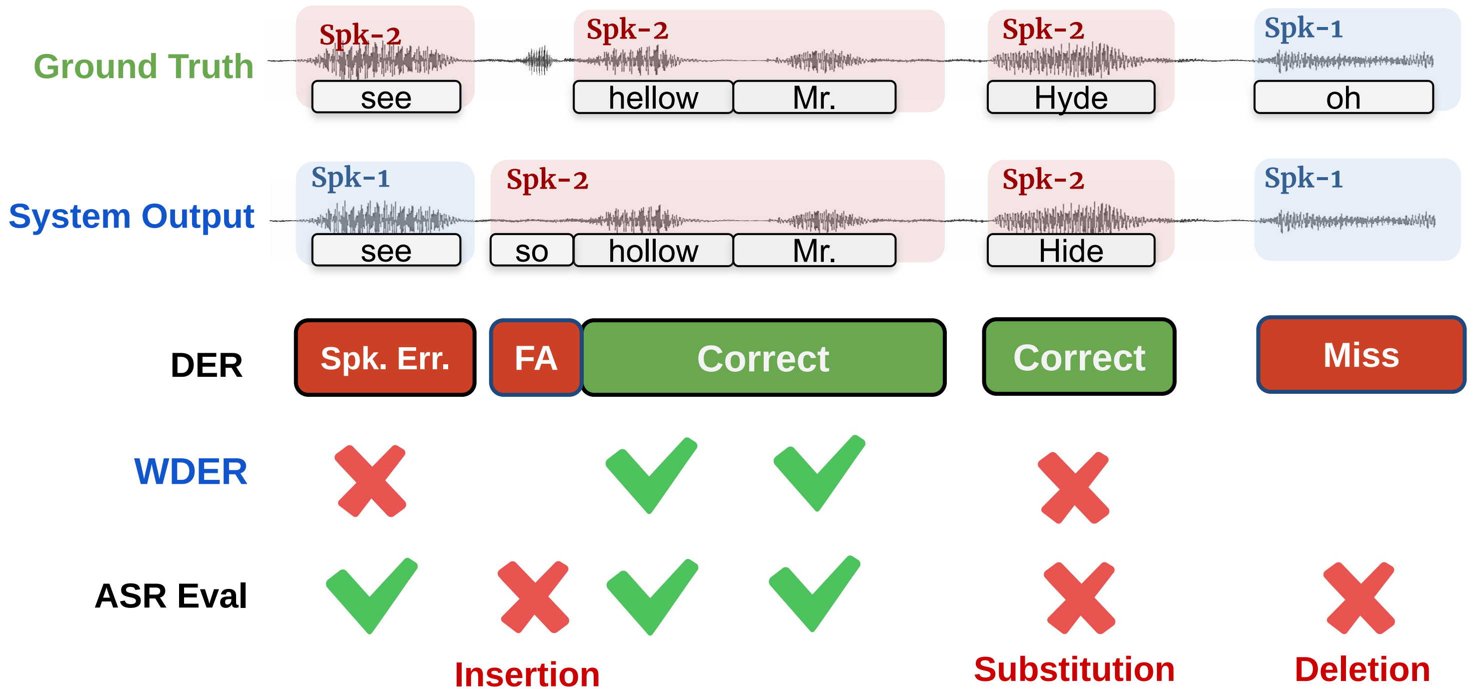


RNN-T structure.

L. Shafey, et al. "Joint speech recognition and speaker diarization via sequence transduction." *Proc. Interspeech*, 2019.

Joint ASR + SD

Word Diarization Error Rate (WDER)



$$WDER = \frac{S_{IS} + C_{IS}}{S + C}$$

S$_{IS}$: # of substitutions with incorrect speaker tokens

C$_{IS}$: # of correct ASR words with incorrect speaker tokens

S: # of substitutions

C: # of correct ASR words

L. Shafey, et al. "Joint speech recognition and speaker diarization via sequence transduction." *Proc. Interspeech*, 2019.

Joint ASR + SD

Word Diarization Error Rate (WDER)

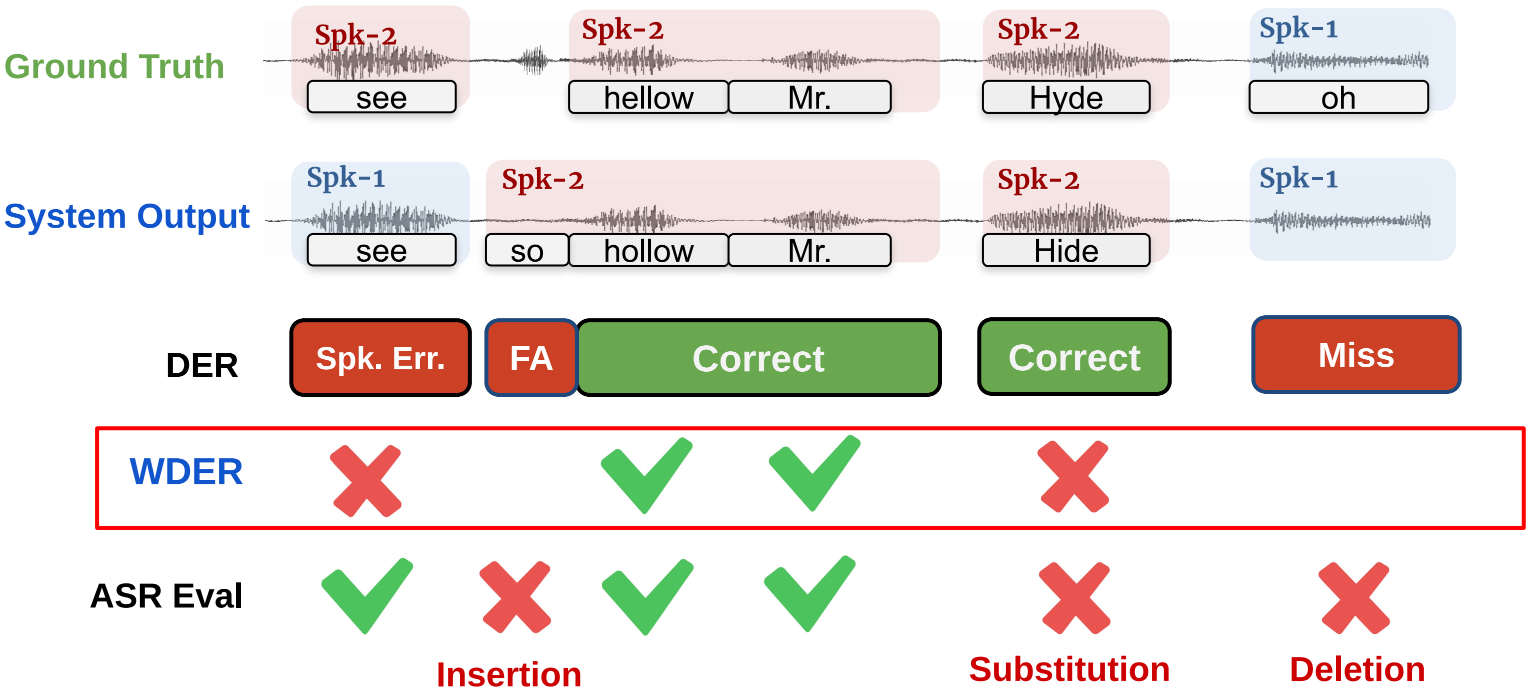
Thoughts on WDER

- Makes sense to consider word level assignment of speaker labels
- Cons: deletion would be encouraged / hard to deal with insertion errors
- Need to consider WER and WDER so they can be supplemental to each other

Yifan Gong (Microsoft)

Joint ASR + SD

Word Diarization Error Rate (WDER)



$$WDER = \frac{S_{IS} + C_{IS}}{S + C}$$

S_{IS}: # of substitutions with incorrect speaker tokens

C_{IS}: # of correct ASR words with incorrect speaker tokens

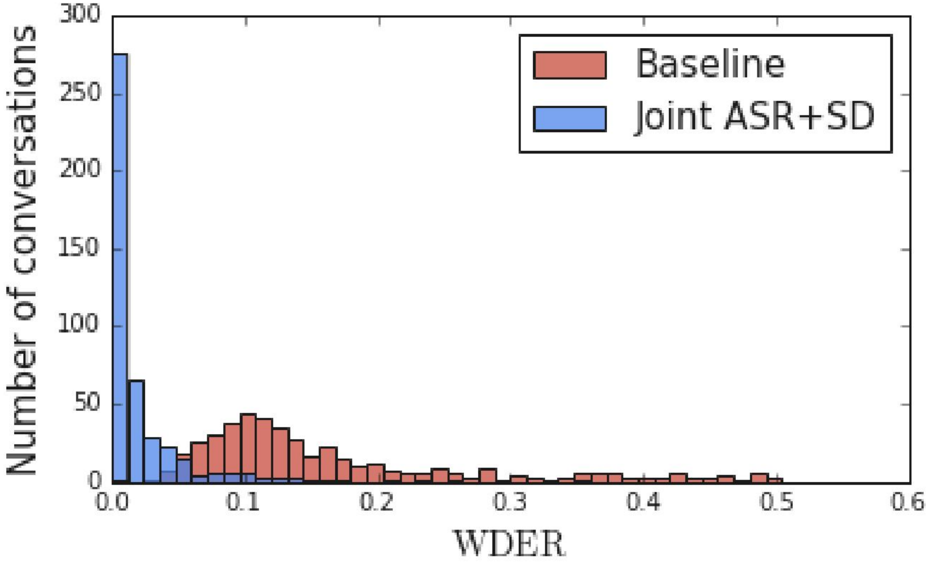
S: # of substitutions

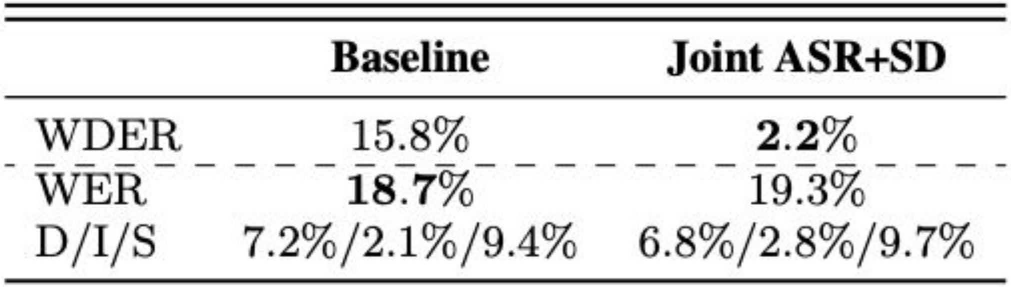
C: # of correct ASR words

L. Shafey, et al. "Joint speech recognition and speaker diarization via sequence transduction." *Proc. Interspeech*, 2019.

Joint ASR + SD

RNN-T for Sequence Transduction of ASR and SD





	Baseline	Joint ASR+SD
WDER	15.8%	**2.2%**
WER	**18.7%**	19.3%
D/I/S	7.2%/2.1%/9.4%	6.8%/2.8%/9.7%

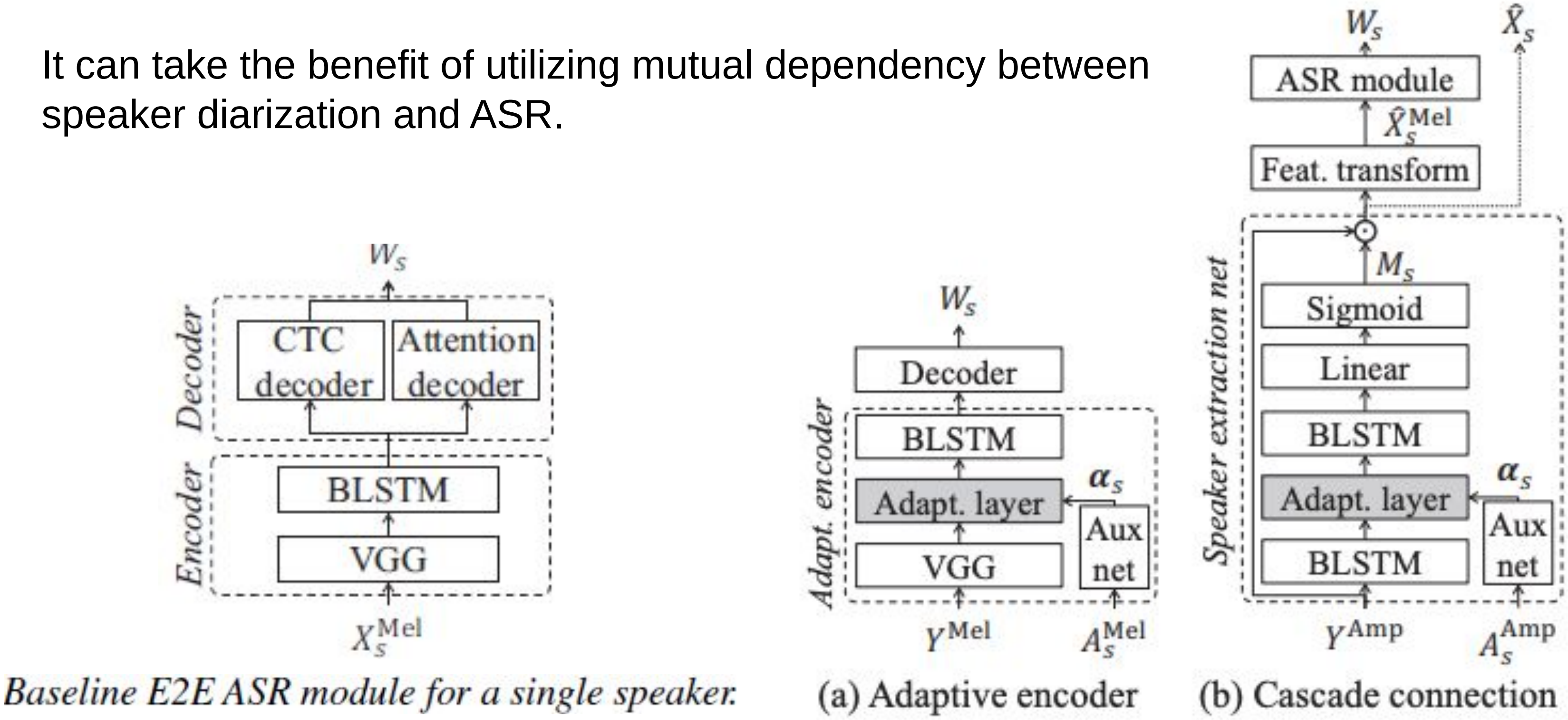
L. Shafey, et al. "Joint speech recognition and speaker diarization via sequence transduction." *Proc. Interspeech*, 2019.

Joint ASR + SD

End-to-end Speaker Beam for Single Channel Target-Speaker ASR

- Jointly models ing approach can be one solution to the decoupling of two systems.

- It can take the benefit of utilizing mutual dependency between speaker diarization and ASR.



Baseline E2E ASR module for a single speaker.

(a) Adaptive encoder

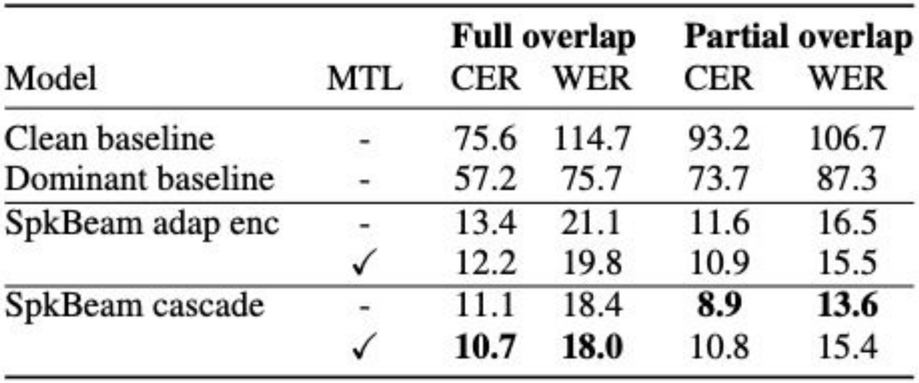
(b) Cascade connection

System architectures of baseline and proposed approach

M. Delcroix, et al., "End-to-end SpeakerBeam for single channel target speech recognition." *Proc. Interspeech*, 2019.

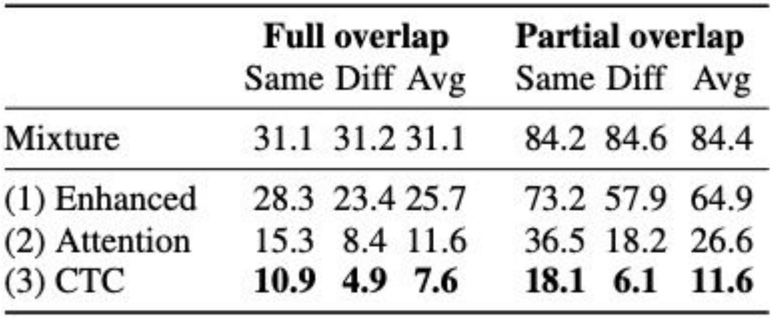
Joint ASR + SD

Results: End-to-end Speaker Beam for Single Channel Target-Speaker ASR



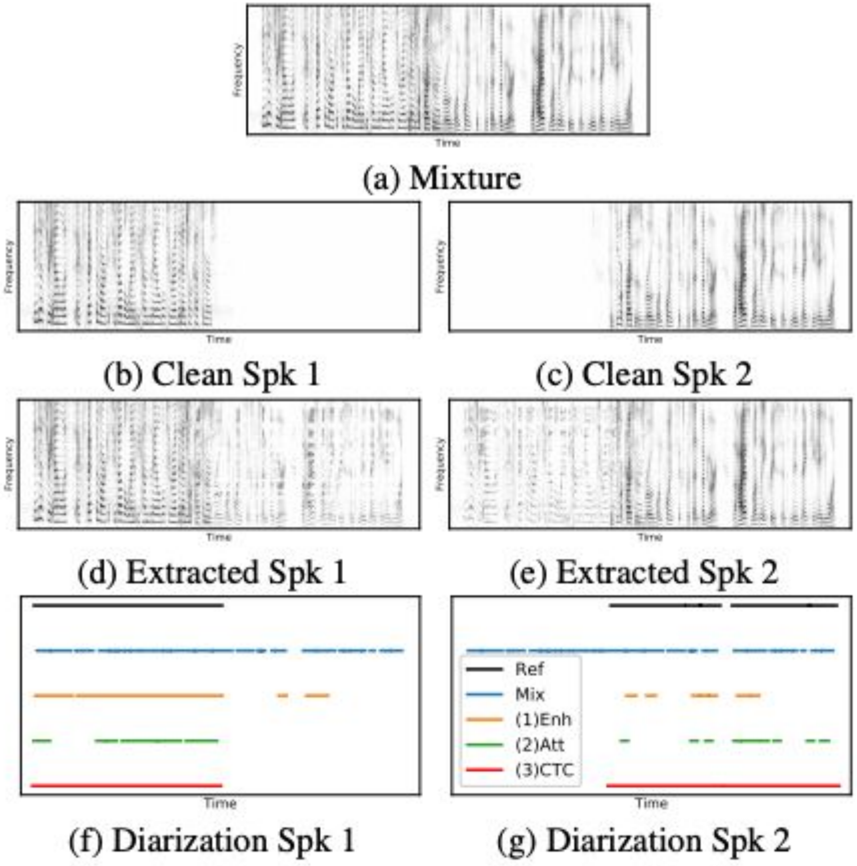
Model	MTL	Full overlap CER	Full overlap WER	Partial overlap CER	Partial overlap WER
Clean baseline	-	75.6	114.7	93.2	106.7
Dominant baseline	-	57.2	75.7	73.7	87.3
SpkBeam adap enc	-	13.4	21.1	11.6	16.5
	✓	12.2	19.8	10.9	15.5
SpkBeam cascade	-	11.1	18.4	**8.9**	**13.6**
	✓	**10.7**	**18.0**	10.8	15.4

Target speech recognition error rates



	Full overlap Same	Full overlap Diff	Full overlap Avg	Partial overlap Same	Partial overlap Diff	Partial overlap Avg
Mixture	31.1	31.2	31.1	84.2	84.6	84.4
(1) Enhanced	28.3	23.4	25.7	73.2	57.9	64.9
(2) Attention	15.3	8.4	11.6	36.5	18.2	26.6
(3) CTC	**10.9**	**4.9**	**7.6**	**18.1**	**6.1**	**11.6**

Diarization error rate



(a) Mixture

(b) Clean Spk 1

(c) Clean Spk 2

(d) Extracted Spk 1

(e) Extracted Spk 2

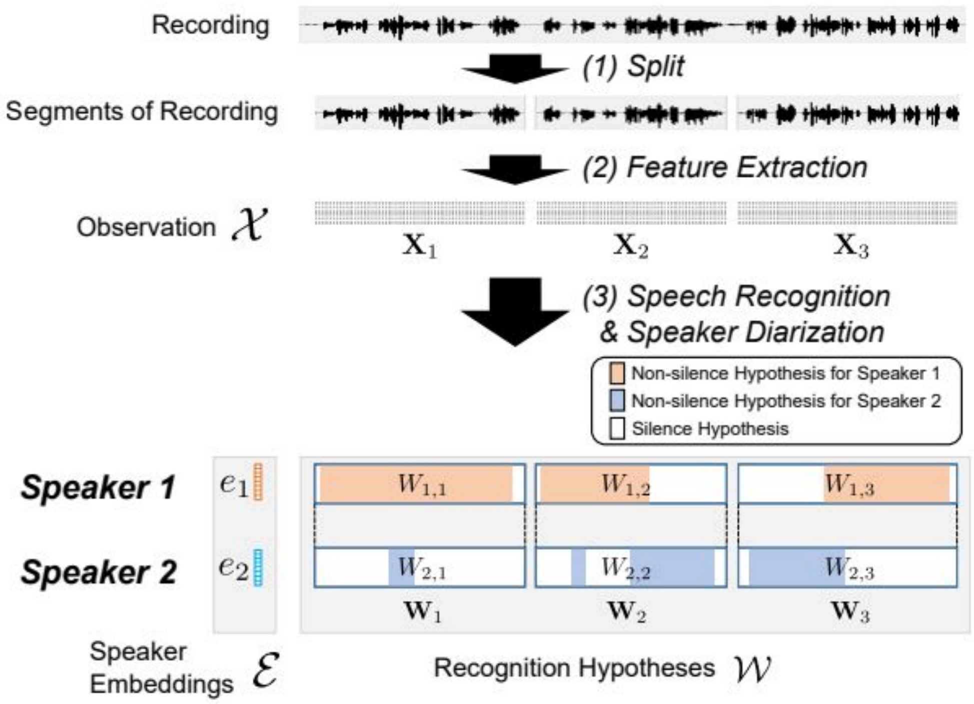
(f) Diarization Spk 1

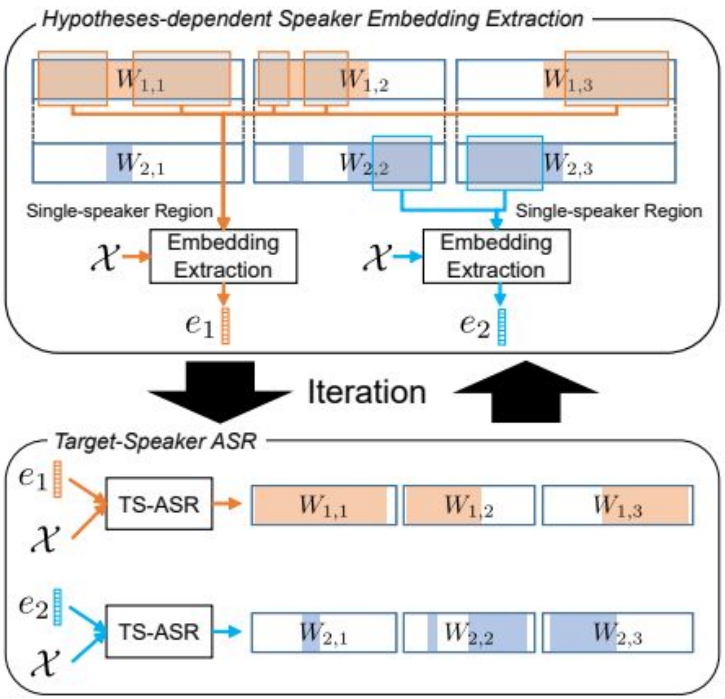
(g) Diarization Spk 2

Examples of speech enhancement and diarization outputs

M. Delcroix, et al., "End-to-end SpeakerBeam for single channel target speech recognition." *Proc. Interspeech*, 2019.

ASR + SD w/ Target-Speaker Acoustic Modeling





Overview of simultaneous ASR and SD

Iterative maximization method between speaker embedding extraction and TS-ASR

N. Kanda, et al., "Simultaneous speech recognition and speaker diarization for monaural dialogue recordings with target-speaker acoustic models." *Proc. ASRU*, 2019.

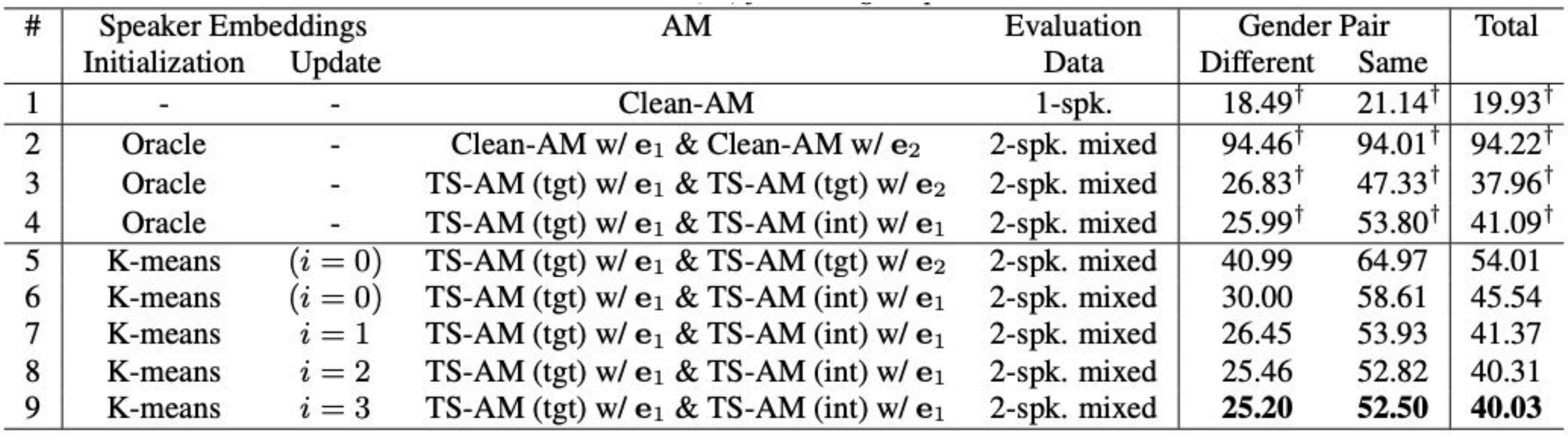
Joint ASR + SD

Maximization of Joint Prob of Speaker Diarization and ASR

Naoyuki Kanda (Microsoft)

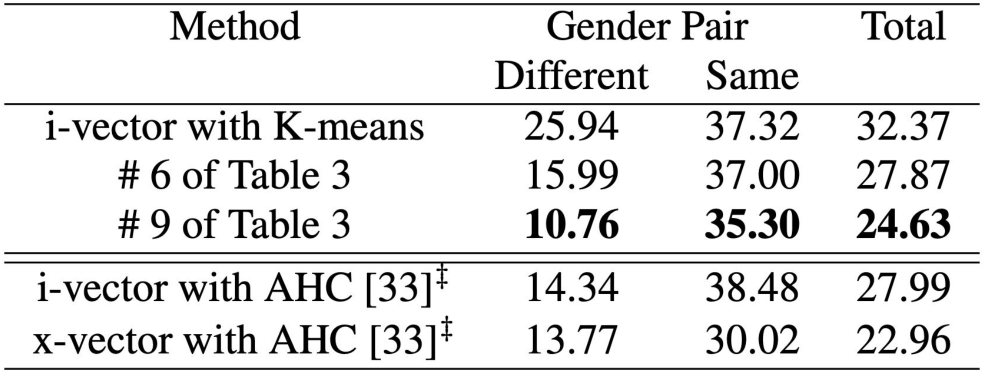
Joint ASR + SD

Results: ASR + SD w/ Target-Speaker Acoustic Modeling



#	Speaker Embeddings		AM	Evaluation	Gender Pair		Total
	Initialization	Update		Data	Different	Same	
1	-	-	Clean-AM	1-spk.	18.49†	21.14†	19.93†
2	Oracle	-	Clean-AM w/ e_1 & Clean-AM w/ e_2	2-spk. mixed	94.46†	94.01†	94.22†
3	Oracle	-	TS-AM (tgt) w/ e_1 & TS-AM (tgt) w/ e_2	2-spk. mixed	26.83†	47.33†	37.96†
4	Oracle	-	TS-AM (tgt) w/ e_1 & TS-AM (int) w/ e_1	2-spk. mixed	25.99†	53.80†	41.09†
5	K-means	$(i = 0)$	TS-AM (tgt) w/ e_1 & TS-AM (tgt) w/ e_2	2-spk. mixed	40.99	64.97	54.01
6	K-means	$(i = 0)$	TS-AM (tgt) w/ e_1 & TS-AM (int) w/ e_1	2-spk. mixed	30.00	58.61	45.54
7	K-means	$i = 1$	TS-AM (tgt) w/ e_1 & TS-AM (int) w/ e_1	2-spk. mixed	26.45	53.93	41.37
8	K-means	$i = 2$	TS-AM (tgt) w/ e_1 & TS-AM (int) w/ e_1	2-spk. mixed	25.46	52.82	40.31
9	K-means	$i = 3$	TS-AM (tgt) w/ e_1 & TS-AM (int) w/ e_1	2-spk. mixed	**25.20**	**52.50**	**40.03**

WERs for dialogue speech



Method	Gender Pair		Total
	Different	Same	
i-vector with K-means	25.94	37.32	32.37
# 6 of Table 3	15.99	37.00	27.87
# 9 of Table 3	**10.76**	**35.30**	**24.63**
i-vector with AHC [33]‡	14.34	38.48	27.99
x-vector with AHC [33]‡	13.77	30.02	22.96

DERs for dialogue speech

N. Kanda, et al., "Simultaneous speech recognition and speaker diarization for monaural dialogue recordings with target-speaker acoustic models." *Proc. ASRU*, 2019.

Chapter 2 Summary

Explore synergies between ASR and Speaker Diarization

In this chapter, we have described

- Early approaches between diarization and ASR
- Use of meta- and linguistic information to facilitate diarization
- Novel e2e approaches for joint ASR and SD

ASR and SD have gone a long way and the technology has matured enough for productization

What will be discussed in Chapter 3

In the wild speaker diarization

- Overlap, short-segment speech
- Domain mismatch
- Inference Speed
- Online Diarization
- Training data for end-to-end system

Chapter 3

Challenges and the State of Speaker Diarization

Chapter 3: Challenges and the State of Speaker Diarization

1. Part 1: Challenges in speaker diarization

1.1. What makes diarization hard?

1.1.1. Overlap speech issues: Chime-6 challenge
1.1.2. Domain mismatch: DIHARD challenge

1.2. Other Challenges

1.2.1. Hurdles for end-to-end diarization system
1.2.2. Inference speed
1.2.3. Online diarization
1.2.4. Segmentation length

2. Part 2: The state of speaker diarization

2.1. Emerging diarization technologies and services

2.1.1. Diarization in conversational AI
2.1.2. Cloud based speech APIs
2.1.3. Diarization with Multi-device/Multi-channel Microphones
2.1.4. Diarization with Better Readability

2.2. The next generation diarization applications

2.2.1. Domain specific applications: healthcare, online video games, social science and security
2.2.2. Diarization for media indexing

Chapter 3

Challenges and the State of Speaker Diarization

Part-1

Challenges in speaker diarization

Challenges in Speaker Diarization: What makes diarization hard?

Ideal Diarization World vs Real Life Diarization World

Diarization is hard!

- humans also have having trouble annotating this challenging diarization dataset.
- far field speech, borderline foreground-background speakers, background music
- Diarization could be even challenging to humans.

Sriram Ganapathy (IISC)

Challenges in Speaker Diarization: What makes diarization hard?

Ideal Diarization World vs Real Life Diarization World

In an ideal world …

- No overlapping speech
- The speech signal is fairly clean
- Limited number of speakers (n < 10)
- Speakers are well distinguishable
- Speaker traits do not vary over time
- Enough domain specific data for diarization

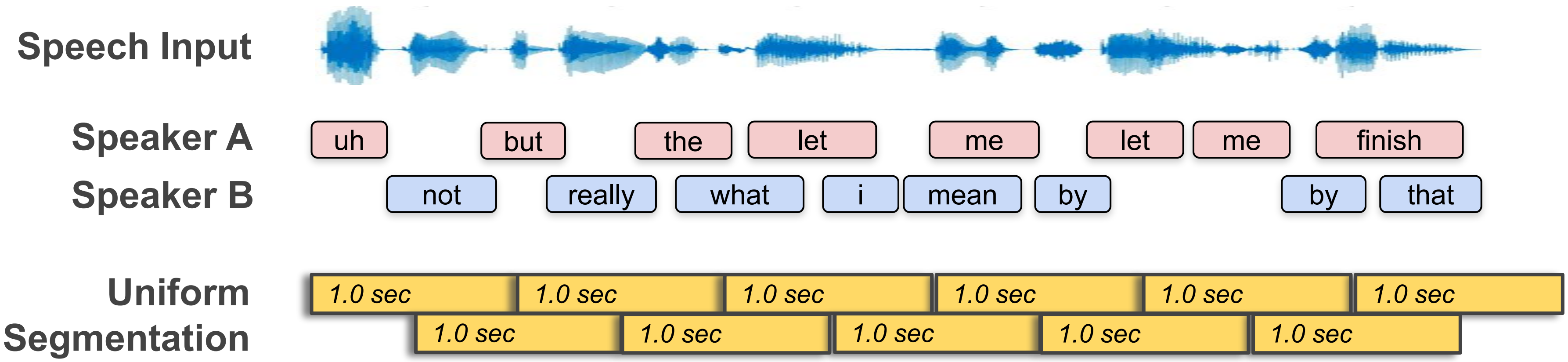
But in real life…

- Overlapping speakers
- Noisy environment
- SAD not working perfectly
- Number of speakers changes over time
- Speakers' traits vary too much
- Segments too short

Challenges in Speaker Diarization: What makes diarization hard?

Overlapping Speech: The killer problem

What if we get multiple speakers in a segment?



- Overlapping speech is very common: In general, overlapping speech occurs 5~15% of total speaking time in two-person dialogue.

- Creates significant amount of DER and loses back channel speech.

Challenges in Speaker Diarization: What makes diarization hard?

Overlap Speech: Killer Problem

Katrin Kirchhoff (Amazon)

Thoughts on overlapping speech

- Overlapping multi-talker speech is a killer problem.
- In some of the worst cases, human listeners have hard time distinguishing the speakers.
- However, In some cases, distinguishing foreground speakers are easily achievable.
- Overlapping speech has lots of potential to be investigated.

Challenges in Speaker Diarization: What makes diarization hard?

Overlap Speech: What is so challenging about overlap speech?

- My work in JSALT workshop was detecting overlap speech and dealing with it.
- Overlap speech can be simulated.
- However, there is a huge gap between simulated overlap and real-life overlap and it makes developing overlap speech detection challenging.

Sriram Ganapathy (IISC)

Challenges in Speaker Diarization: What makes diarization hard?

Overlapping Speech 3

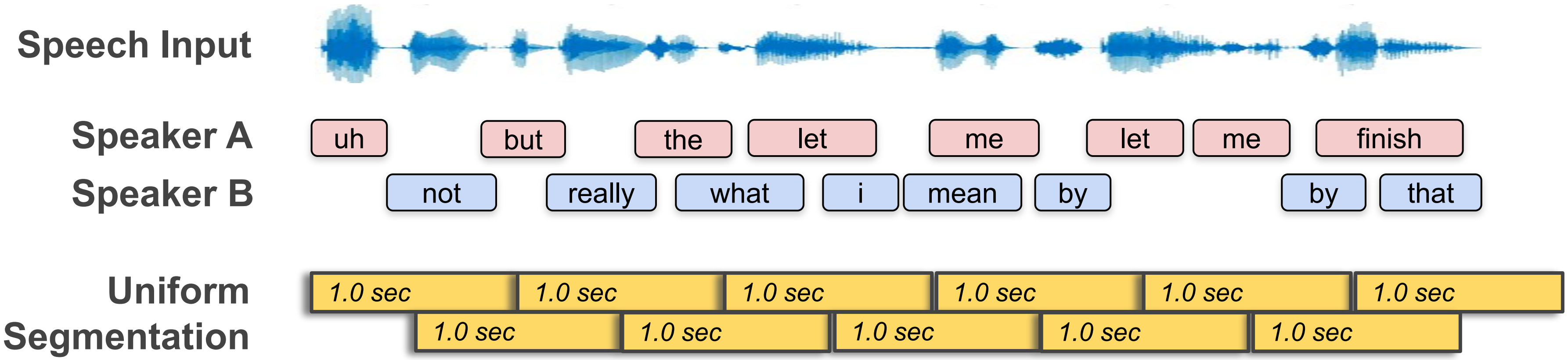
Yifan Gong (Microsoft)

Thoughts on Overlap Speech

- Even human speakers ask to “say it again” when overlap speech happens
- Machines have better chance to deal with overlap speech in the future.

Challenges in Speaker Diarization: What makes diarization hard?

Overlapping Speech: The killer problem

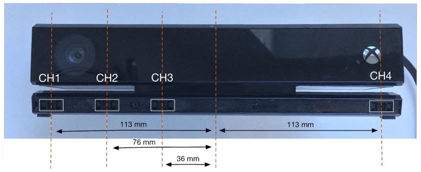


- Solutions for overlapping speech:
  - Overlap detection and assign system
  - Resegmentation
  - Target-Speaker Voice Activity Detection
  - Speech Separation

Challenges in Speaker Diarization: What makes diarization hard?

Chime Challenge

“The problem of distant multi-microphone conversational speech **diarization and recognition** in everyday home environments”



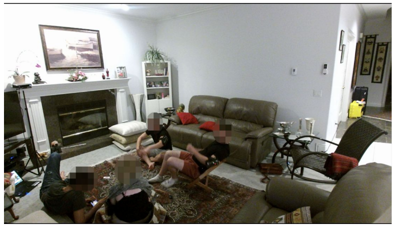
For challenge
4-ch Kinect Microphone Array

For transcription
Personal Binaural Microphones
(Worn by participants)

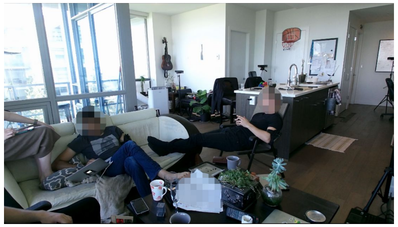
Location

Kitchen

Dining



Living Room 1



Living Room 2

Scenario

- Twenty separate dinner parties that are taking place in **real homes**
- Each dinner party has **four participants.**
- **Realistic and in the wild conversation with lots of overlap and back channel speech**

https://chimechallenge.github.io/chime6

Chime Challenge

Evaluation Condition*:

- An accurate array synchronization script was provided,
- **the impact of diarization error** on speech recognition error was measured,
- upgraded, state-of-the-art baselines are provided for diarization, enhancement, and recognition.

*Includes some portion of DIHARD challenge dataset

6th CHiME Speech Separation and Recognition Challenge (CHiME-6) result release at ICASSP 2020

- **Track 1:** Multiple-array speech recognition **(ASR only)**
- **Track 2:** Multiple-array diariazation and recognition **(Diarization + ASR)**

https://chimechallenge.github.io/chime6

Challenges in Speaker Diarization: What makes diarization hard?

Overlapping Speech - Chime challenge and diarization

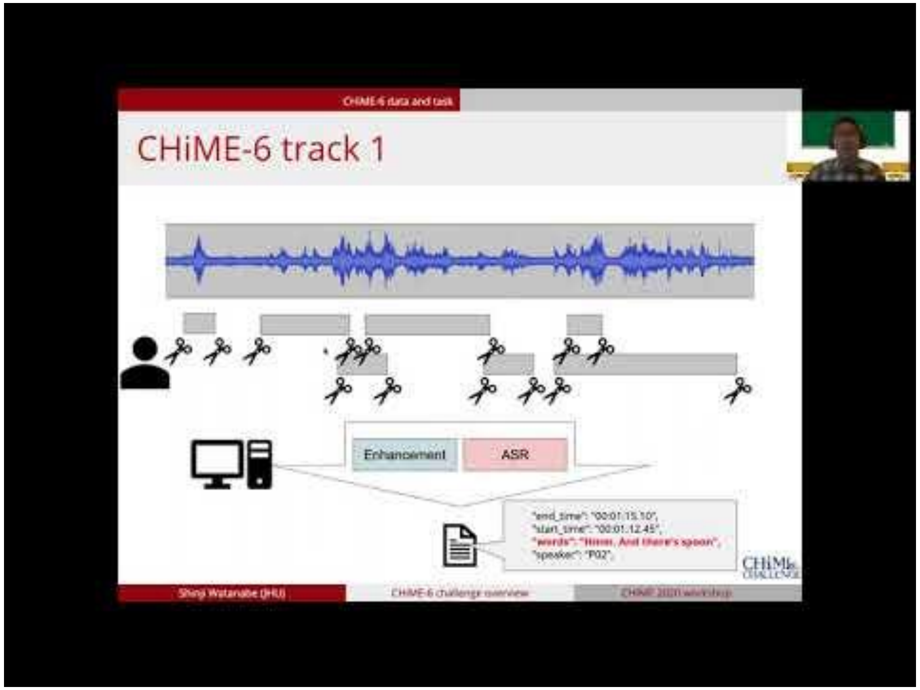
Shinji Watanabe (JHU)

Thoughts on overlapping speech

- In CHIME 5 Challenge, speaker labels are given to ASR module assuming that diarization is already done perfectly.
- CHIME 6 track 1 is equal to CHIME 5.
- Having the oracle diarization result could not be realistic enough.
- We are thinking about including diarization to the upcoming CHIME challenges.

Challenges in Speaker Diarization: What makes diarization hard?

Overlapping Speech - Chime challenge and diarization



CHiME-6 data and task

CHiME-6 track 1

"end_time": "00:01:15.10",
"start_time": "00:01:12.45",
"words": "Hmm. And there's spoon",
"speaker": "P02",

Shinji Watanabe (JHU) | CHiME-6 challenge overview | CHiME 2020 workshop

CHIME-6 data example

- Lots of overlapping speech
- Background/environmental/recording device noise
- Conversational speech
- Distant microphones

Challenges in Speaker Diarization: What makes diarization hard?

Overlapping Speech - Chime challenge

Paola Garcia (JHU)

Diarization and ASR result

- In CHIME 6 track 2, oracle diarization result is not provided.
- Multiple microphones are employed in CHIME challenge .
- We Combined SAD outputs and PLDA results.
- We used 0.25 second of window hop-length and performed overlap assignment with the results.
- We got really good diarization result but it did not improve ASR WER result.

Challenges in Speaker Diarization: What makes diarization hard?

Overlapping Speech - Chime challenge

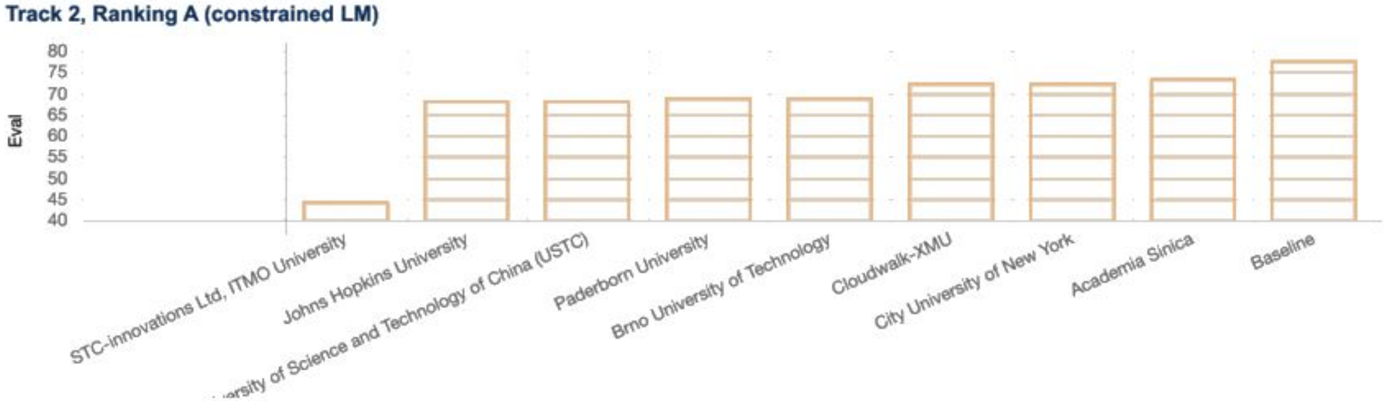
Naoyuki Kanda (Microsoft)

Chime Challenge Takeaways for Overlap Speech:

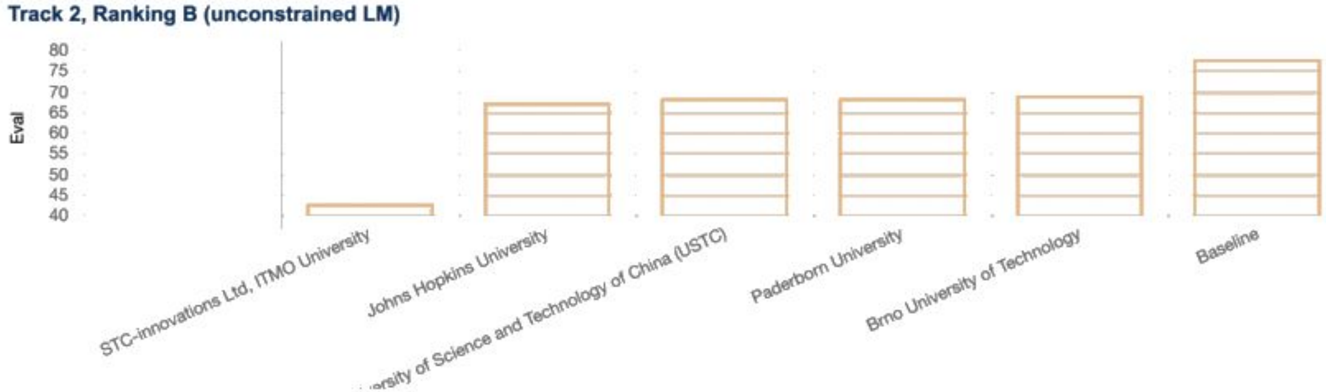
- STC team's target speaker VAD showed superior performance.
- Guided source separation with speaker diarization if diarization result is good.
- STC team showed that the possibility of using the combination of target speaker VAD and diarization to obtain superior diarization performance.

Challenges in Speaker Diarization: What makes diarization hard?

CHIME-6 Track 2 (Diar+ASR) Winner: STC


Track 2, Ranking A (constrained LM)

Track 2 (constrained LM), best performing system (STC) WER: Dev: 41.6 %, Eval 44.5 %

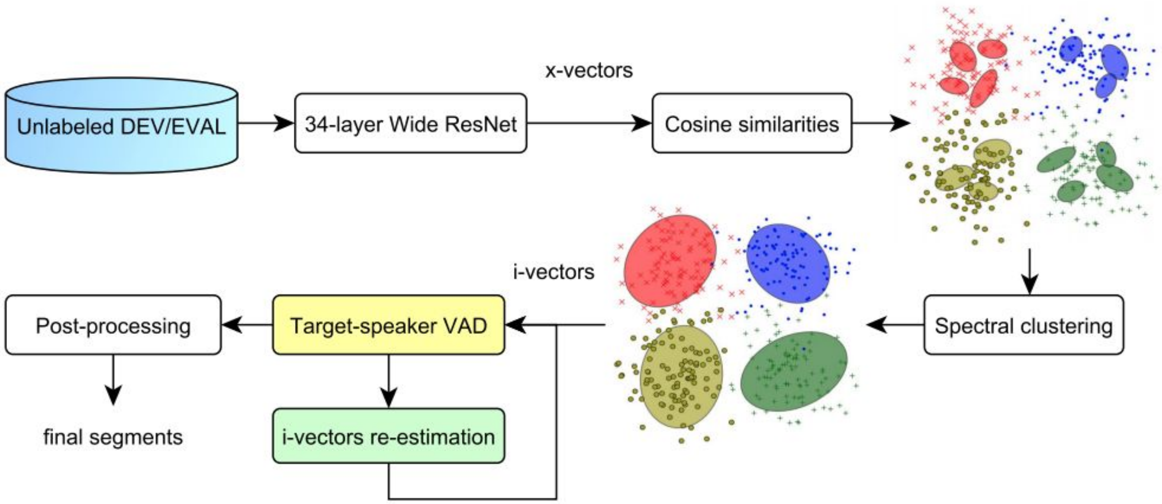

Track 2, Ranking B (unconstrained LM)

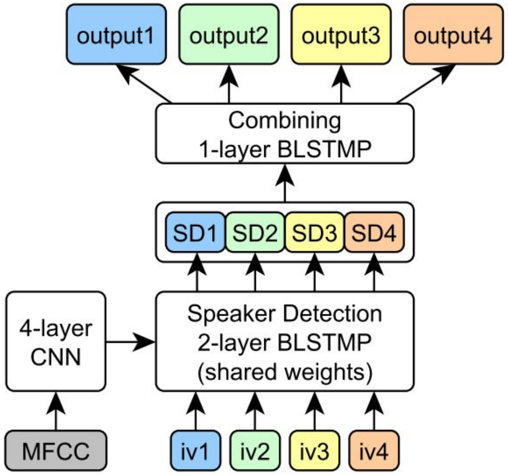
Track 2 (unconstrained LM), best performing system (STC) WER: Dev: 39.6 % Eval 42.7 %

https://chimechallenge.github.io/chime6/results.html

Challenges in Speaker Diarization: What makes diarization hard?

CHIME-6 Track 2 (Diar+ASR) Winner: STC system [1]



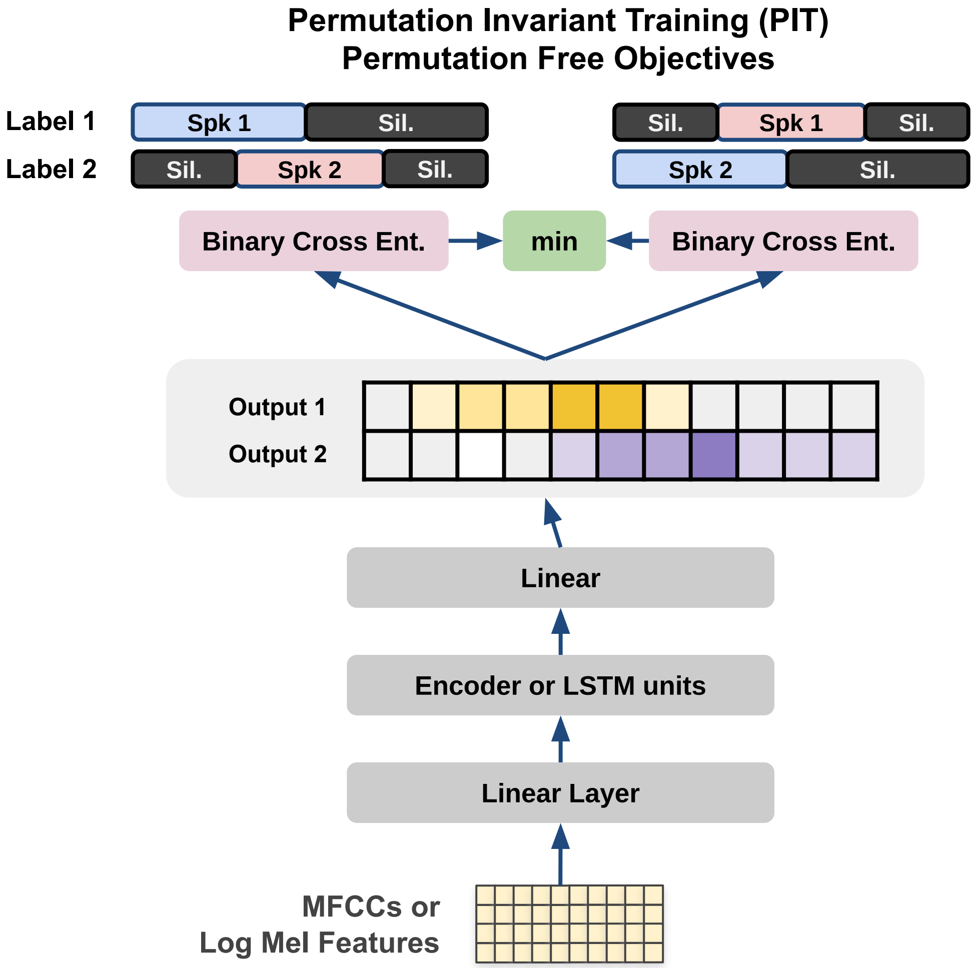


- ResNet inspired x-vectors
- Cosine Similarities with Auto-tuning Spectral Clustering method (NME-SC[2])
- Target-speaker VAD (TS-VAD) greatly improved the overall performance
  - Uses i-vector input from parallel streams of speaker detection (SD) blocks
  - STC's TS-VAD shows that target-speaker VAD can be a solution for overlapping speech

[1] https://chimechallenge.github.io/chime2020-workshop/papers/CHiME_2020_paper_medennikov.pdf
[2] Taejin Park et. al. "Auto-Tuning Spectral Clustering for Speaker Diarization Using Normalized Maximum Eigengap" IEEE SPL. 2019, p.381-385.

E2E Neural Speaker Diarization (EEND) with Permutation-Free Objectives

- Inspired by sound event detection handling multi-label classification
- Permutation-free scheme introduced to figure out the permutation problem
- Both deals with overlapping speech as well as minimizing diarization errors



Fujita, Yusuke, et al. "End-to-end neural speaker diarization with permutation-free objectives." arXiv preprint arXiv:1909.05952 (2019).

Challenges in Speaker Diarization: What makes diarization hard?

E2E Neural Speaker Diarization (EEND) with Permutation-Free Objectives

Shinji Watanabe (JHU)

Permutation Invariant Training (PIT) and source separation for End-to-end speaker diarization

- This idea came from audio event detection and source separation.
- We are inspired by permutation problem from DCASE challenge (audio event detection challenge).
- We are also inspired by speech separation where permutation invariant training (PIT) is needed.

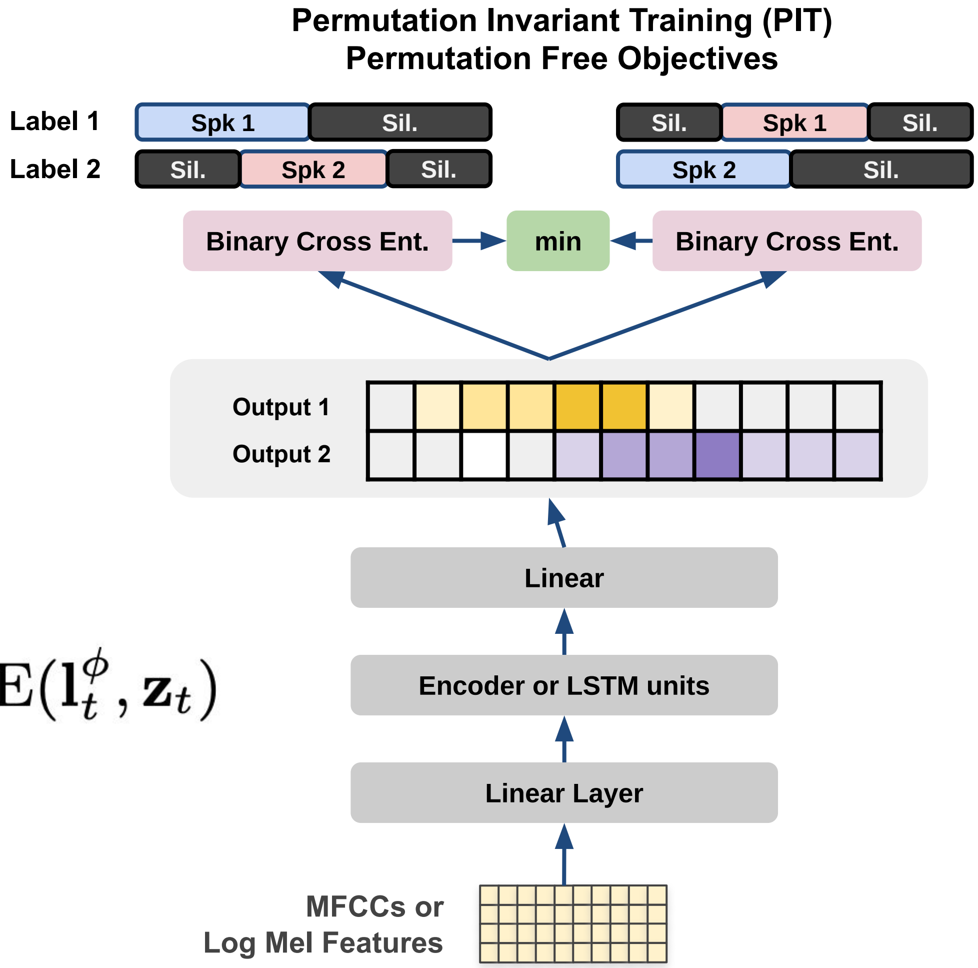
https://arxiv.org/pdf/1909.05952.pdf

Challenges in Speaker Diarization: What makes diarization hard?

E2E Neural Speaker Diarization (EEND) with Permutation-Free Objectives

- Inspired by sound event detection handling multi-label classification
- Permutation-free scheme introduced to figure out the permutation problem
- Both deals with overlapping speech as well as minimizing diarization errors

$$J^{\text{PIT}} = \frac{1}{TC} \min_{\phi \in \text{perm}(C)} \sum_t \text{BCE}(\mathbf{l}_t^{\phi}, \mathbf{z}_t)$$



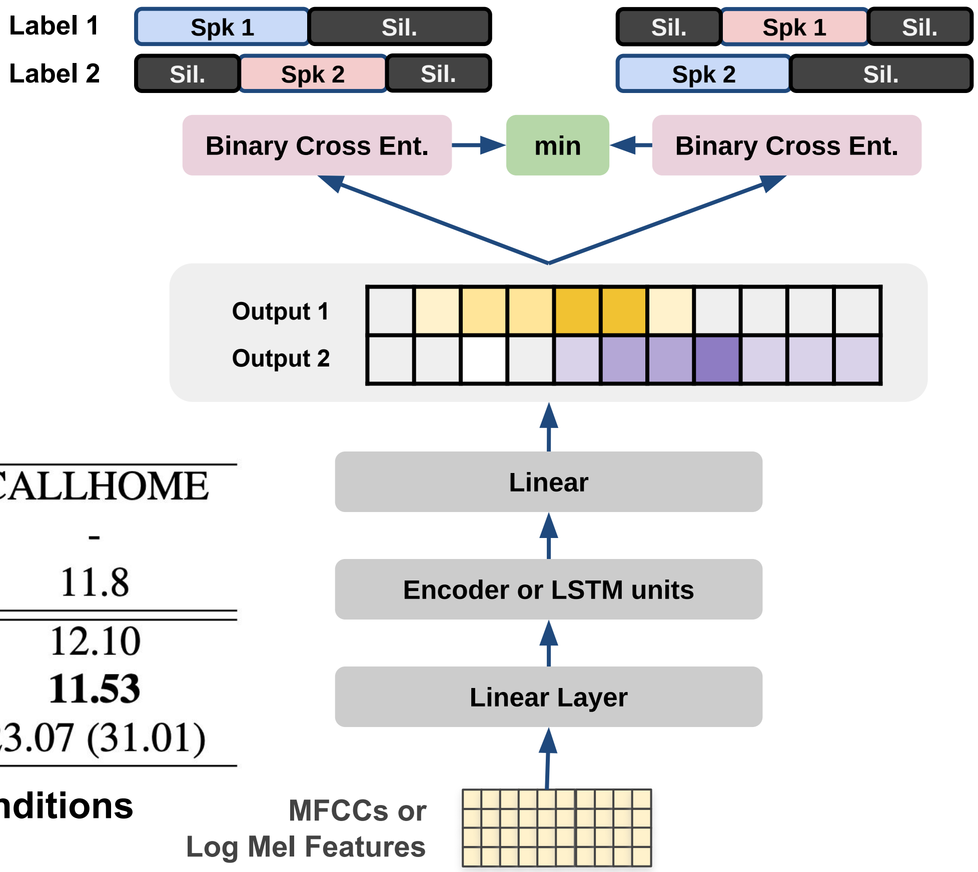
Fujita, Yusuke, et al. "End-to-end neural speaker diarization with permutation-free objectives." arXiv preprint arXiv:1909.05952 (2019).

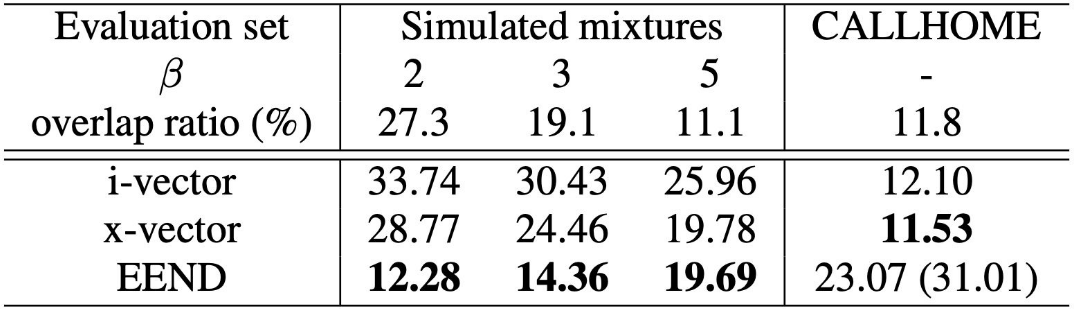
Challenges in Speaker Diarization: What makes diarization hard?

E2E Neural Speaker Diarization (EEND) with Permutation-Free Objectives

- Inspired by sound event detection handling multi-label classification
- Permutation-free scheme introduced to figure out the permutation problem
- Both deals with overlapping speech as well as minimizing diarization errors

Permutation Invariant Training (PIT)
Permutation Free Objectives





Evaluation set	Simulated mixtures			CALLHOME
β	2	3	5	-
overlap ratio (%)	27.3	19.1	11.1	11.8
i-vector	33.74	30.43	25.96	12.10
x-vector	28.77	24.46	19.78	**11.53**
EEND	**12.28**	**14.36**	**19.69**	23.07 (31.01)

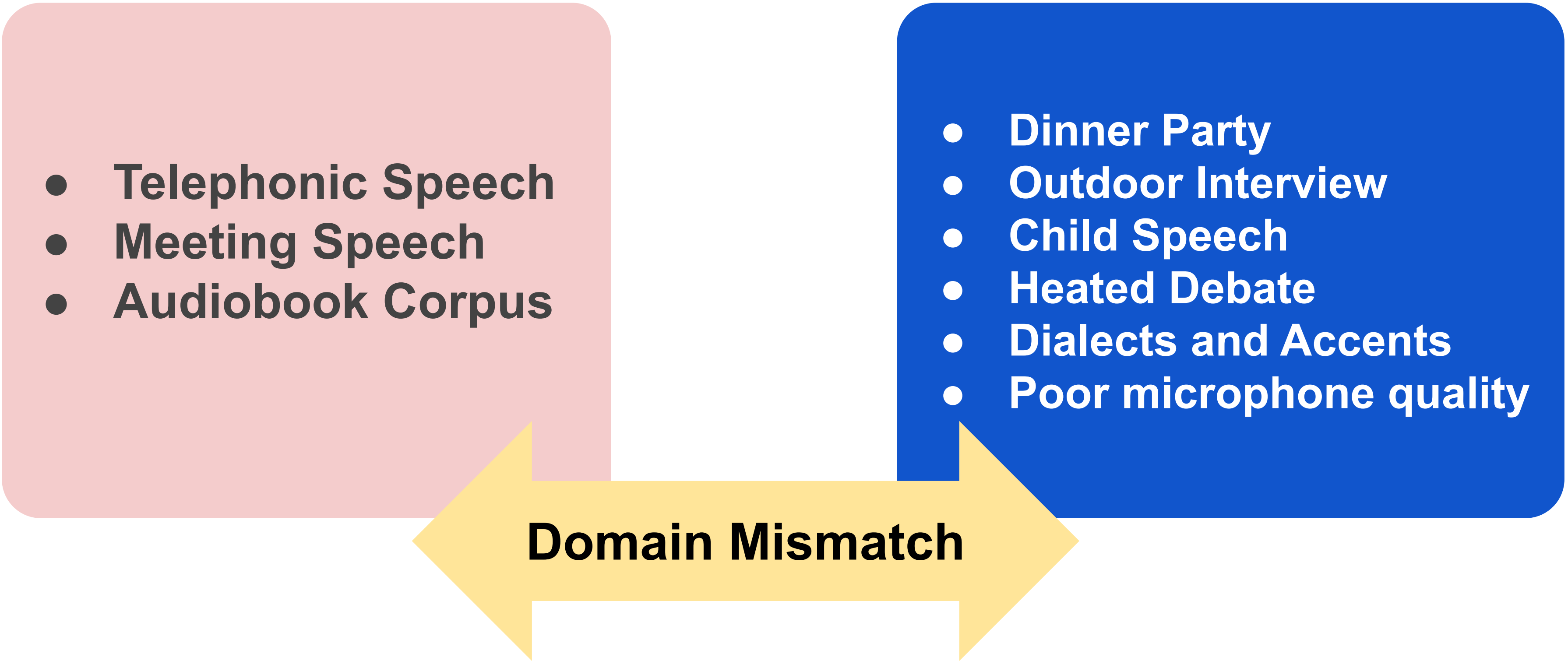
DERs rates on different overlapping conditions

Fujita, Yusuke, et al. "End-to-end neural speaker diarization with permutation-free objectives." arXiv preprint arXiv:1909.05952 (2019).

Challenges in Speaker Diarization: What makes diarization hard?

Domain Mismatch

What we have for training | **In the wild conditions**



- Telephonic Speech
- Meeting Speech
- Audiobook Corpus

- Dinner Party
- Outdoor Interview
- Child Speech
- Heated Debate
- Dialects and Accents
- Poor microphone quality

Challenges in Speaker Diarization: What makes diarization hard?

Domain Mismatch 2

Douglas Reynolds
(MIT Lincoln Lab)

Domain mismatch

- Domain mismatch has been the primary limiter in speaker-ID problems.
- Diarization brings another twist where we see behavioral shift and temporal shift.
- In diarization, there is a temporal aspect of how people interact.
- For example, broadcast news one person speaks for a long time.
- In meetings, one person dominates or people talk back and forth.
- This creates all kinds of temporal dynamics and makes diarization hard to model.

Challenges in Speaker Diarization: What makes diarization hard?

Domain Mismatch - How is child speech different?

Paola Garcia (JHU)

Child speech domain

- Child speech is completely wild.
- Kids are not collaborative and usually show unexpected behavior.
- Sometimes kids do not want to answer and stay silent.
- We should keep in mind Indoor and outdoor scenarios due to the nature of interview.
- Nearly all of our systems failed dramatically on child speech domain.

Challenges in Speaker Diarization: What makes diarization hard?

Domain Mismatch -

Domain mismatch

- Domain mismatch is one of the major problems in speech modeling.
- There is a gap between simulated environment and real-life environment.

Andreas Stolcke (Amazon)

Challenges in Speaker Diarization: What makes diarization hard?

Domain Mismatch

Challenging mismatch problems in diarization

- Intra-speaker variability: same speakers sound differently even within a session or between sessions

- Audio context: the location and situation where the audio even is happening

Shri Narayanan (USC)

Challenges in Speaker Diarization: What makes diarization hard?

Domain Mismatch - DIHARD-2 Challenge Review

Diarization is Hard: Strictly evaluated diarization on challenging domains

Diarization Evaluation in DIHARD 2:

- Evaluate the overlapping regions.
- No 0.25s of collar when the output is evaluated
- JER (Jaccard Error Rate) is employed

Tracks:

- **Track 1:** Oracle SAD + Single channel Diarization
- **Track 2:** System SAD + Single channel Diarization
- **Track 3:** Oracle SAD + Multi channel Diarization
- **Track 4:** System SAD + Multi channel Diarization

Dataset Domains:

- Audiobooks:
- Broadcast interview
- Child language (6-18 month old)
- Clinical (12-16 old children)
- Court room
- Map task
- Meeting
- Restaurant
- Sociolinguistic field recordings
- sociolinguistic lab meetings
- web video

Challenges in Speaker Diarization: What makes diarization hard?

Domain Mismatch 2

Sriram Ganapathy (IISC)

What was the motivation of DIHARD challenge?

- DIHARD challenge started at JSALT workshop in 2017
- While we were building baselines for diarization systems, we realized that diarization systems are very domain specific.
- We were motivated to create a evaluation set which people can test their diarization system for many different challenging domains
- DIHARD evaluation pursues domain-agnostic diarization system that can work on lots of different domains.

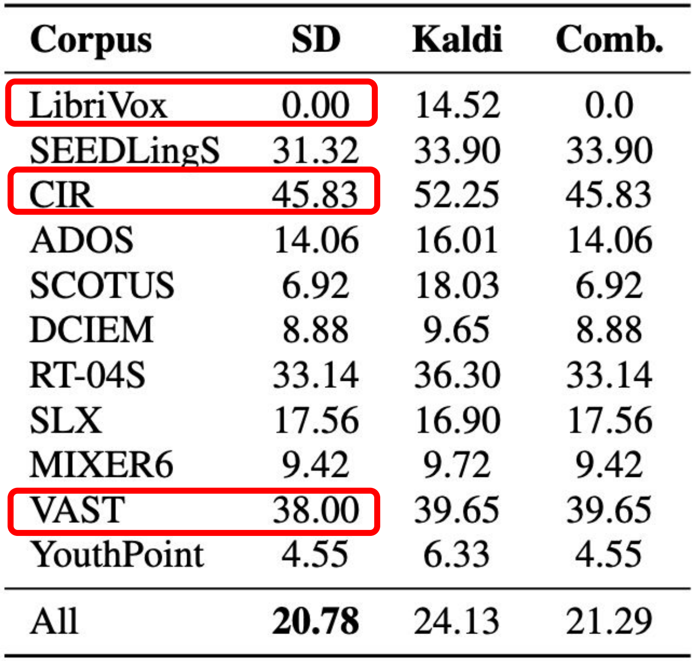
Challenges in Speaker Diarization: What makes diarization hard?

Domain Mismatch - DIHARD2

- **Domain mismatch creates huge error in challenging diarization tasks.**

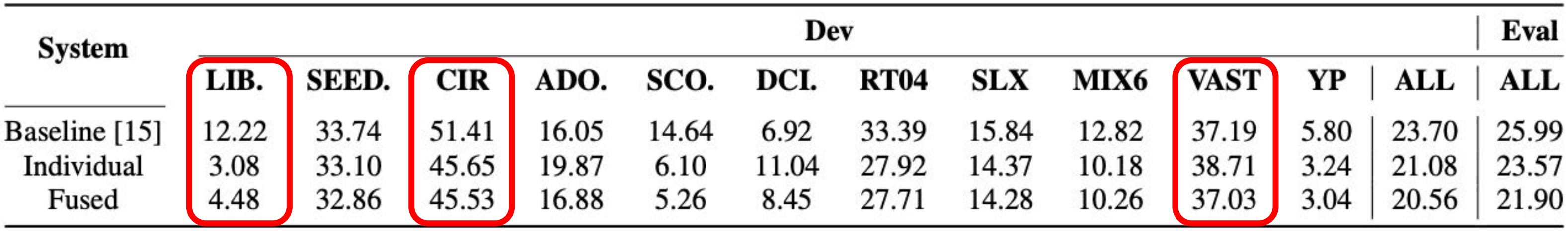
LibriVox: Audiobooks (1 spk/sess)
SEEDLingS: Child language (3.6 spks/sess)
ADOS: Clinical (2.1spk /sess)
SCOTUS: Court room (6.9 spk/sess)
DCIEM: Map task (2 spk/sess)
ROAR: Meeting (3.9 spk/sess)
CIR: Restaurant (6.4 spk/sess)
MIXER6: Sociolinguistic field recordings (2spk/sess)
SCO: sociolinguistic lab meetings (7.3spk/sess)
SLX: sociolinguistic interviews (3.5 spk/sess)
VAST: web video (3.5 spk /sess)

UWB-NTIS's system results



Corpus	SD	Kaldi	Comb.
LibriVox	0.00	14.52	0.0
SEEDLingS	31.32	33.90	33.90
CIR	45.83	52.25	45.83
ADOS	14.06	16.01	14.06
SCOTUS	6.92	18.03	6.92
DCIEM	8.88	9.65	8.88
RT-04S	33.14	36.30	33.14
SLX	17.56	16.90	17.56
MIXER6	9.42	9.72	9.42
VAST	38.00	39.65	39.65
YouthPoint	4.55	6.33	4.55
All	**20.78**	24.13	21.29

LEAP's system results



System	Dev												Eval
	LIB.	SEED.	CIR	ADO.	SCO.	DCI.	RT04	SLX	MIX6	VAST	YP	ALL	ALL
Baseline [15]	12.22	33.74	51.41	16.05	14.64	6.92	33.39	15.84	12.82	37.19	5.80	23.70	25.99
Individual	3.08	33.10	45.65	19.87	6.10	11.04	27.92	14.37	10.18	38.71	3.24	21.08	23.57
Fused	4.48	32.86	45.53	16.88	5.26	8.45	27.71	14.28	10.26	37.03	3.04	20.56	21.90

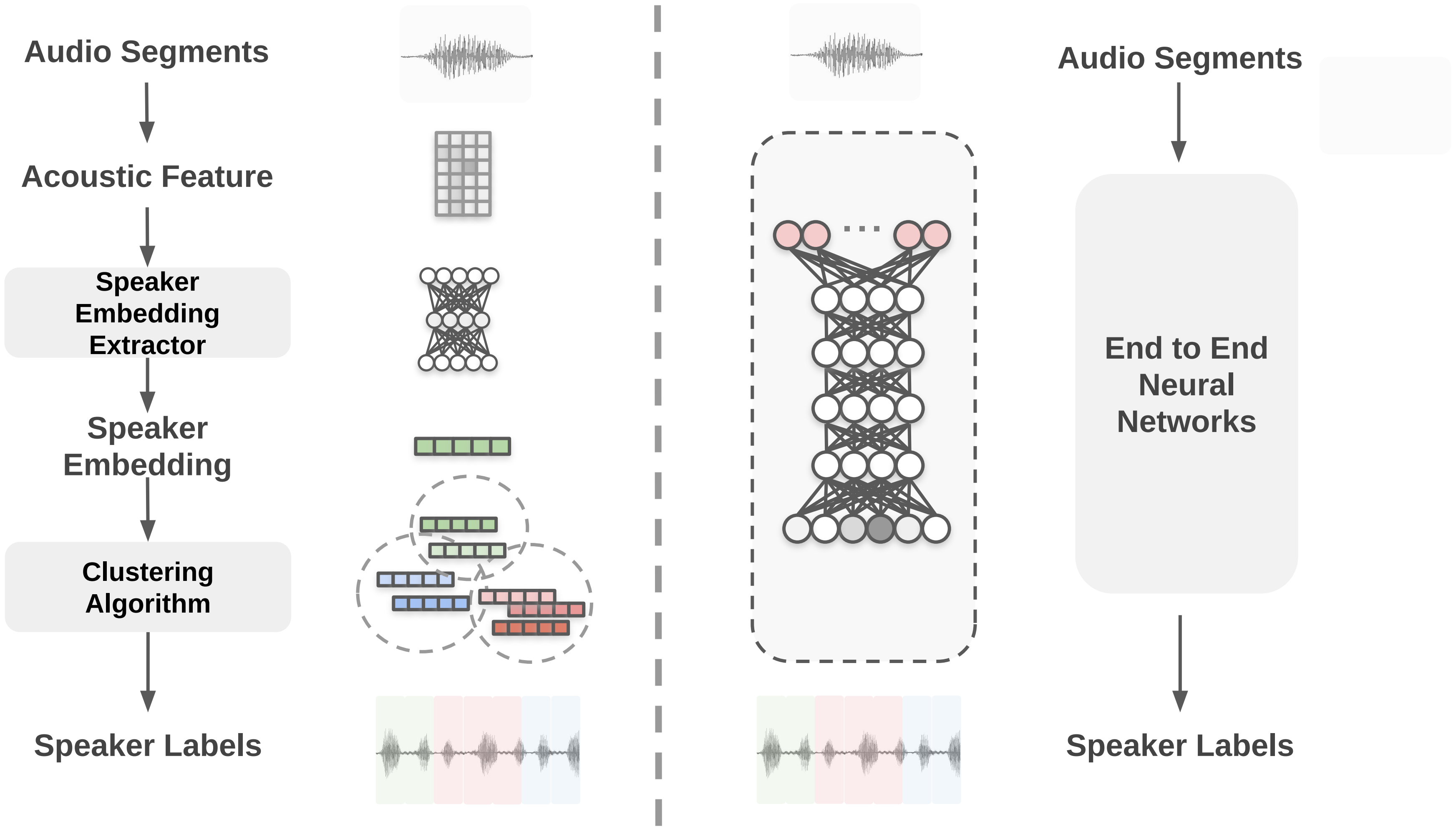
Singh, Prachi, et al. "LEAP diarization system for the second dihard challenge." (2019).
Zajíc, Zbyněk, et al. "UWB-NTIS speaker diarization system for the DIHARD II 2019 challenge." arXiv preprint arXiv:1905.11276 (2019).

Challenges in Speaker Diarization: What makes diarization hard?

End-to-End Diarization and Training Datasets

Modular Diarization

End2End Diarization



Audio Segments

Acoustic Feature

Speaker Embedding Extractor

Speaker Embedding

Clustering Algorithm

Speaker Labels

Audio Segments

End to End Neural Networks

Speaker Labels

Challenges in Speaker Diarization: What makes diarization hard?

End-to-End Diarization and Training Datasets

Thoughts on end-to-end diarization model?

- The definition of end-to-end model: A model that is optimized by one function.
- I believe that diarization system is better to be optimized in a single model.
- End-to-end approaches are now common in other fields such as ASR.

Shinji Watanabe (JHU)

Challenges in Speaker Diarization: What makes diarization hard?

End-to-End Diarization and Training Datasets

Quan Wang (Google)

Thoughts on End-to-end Speaker Diarization

- End-to-end systems seem very promising and look positive.
- however, at the moment, End-to-end systems seem to be in beta state.
- We need high quality data and no such dataset yet exists.
- Until we have high quality and sizable diarization datasets, modular diarization can still be employed.

Challenges in Speaker Diarization: What makes diarization hard?

End-to-End Diarization and Training Datasets

Shinji Watanabe (JHU)

Downside of end-to-end diarization system

- Label problem: not consistent across over the datasets or applications

Challenges in Speaker Diarization: What makes diarization hard?

End-to-End Diarization and Training Datasets

Modular Diarization End2End Diarization

	Modular Diarization	End2End Diarization
SoTA (April 2020) on CallHome Dataset	[1]Spk. Err 5~6% (System SAD) [1]DER 6~7% (Oracle SAD)	[2]Spk. Err > 10% (System SAD)
Training Data	Relatively **easy** to get (Separately train each module: embedding, clustering, language model)	Relatively **hard** to get üNumber of speakers üAcoustic environment üLanguage
Training Steps	Relatively **complicated**	Relatively **simple**
Validation of Each Function	Relatively **easy** (Separately test segmentation, embedding and clustering)	Relatively **hard**
Proper Applications	Media indexing Offline dialogue analysis	Online ASR pipeline Real-time dialogue system

[1] Fujita, Yusuke, et al. "End-to-End Neural Speaker Diarization with Self-attention." *arXiv preprint arXiv:1909.06247*, 2019

[2] Lin, Qingjian, et al. "LSTM based Similarity Measurement with Spectral Clustering for Speaker Diarization." Interspeech 2019

Other Challenges

Inference speed of speaker diarization system

- As diarization systems get improved, the inference speed become slower.
- Iterative approaches make speaker diarization system very slow
- In real life scenario, the slow inference of diarization output gives rise to practical problems.
- Not only the speed, the resource for the inference (Heavy CPU/GPU usage)

Xavier Anguera (ELSA)

Challenges in Speaker Diarization

Other Challenges

Online diarization

Miguel Jette (Rev.ai)

Andreas Stolcke (Amazon)

Challenges in Speaker Diarization

Other Challenges

Segment length

- Fixed window length segmentation gives a limitation of fixed output resolution.
- Diarization systems in the future needs to have variable window length to give more flexibility.
- A strategy that can fuse the scores from multiple scales to increase the temporal resolution is needed.

Sriram Ganapathy (IISC)

Other Challenges

Neural Net Regime w/ No Signal Understanding

- Neural nets working great
- However, more understanding on speech signals would be also required
- Signal processing minds + computer science would be a great combination to address problems

Xavier Anguera (ELSA)

Other Challenges

Interpretability

- Hard to answer customers questioning why bad diarization results
- Explainability of what caused errors, very important to customers

Miguel Jette (Rev.ai)

Chapter 3 Part 1 Summary

Main Challenges

- Overlapping speech
    - CHIME-6 Track-2
    - Permutation invariant training
- Domain mismatch
    - DIHARD

Other Challenges

- Data problems for end-to-end speaker diarization
- Inference speed
- Online diarization
- Fine-grained resolution for embedding processing
- Neural network regime w/o signal understanding
- Interpretability

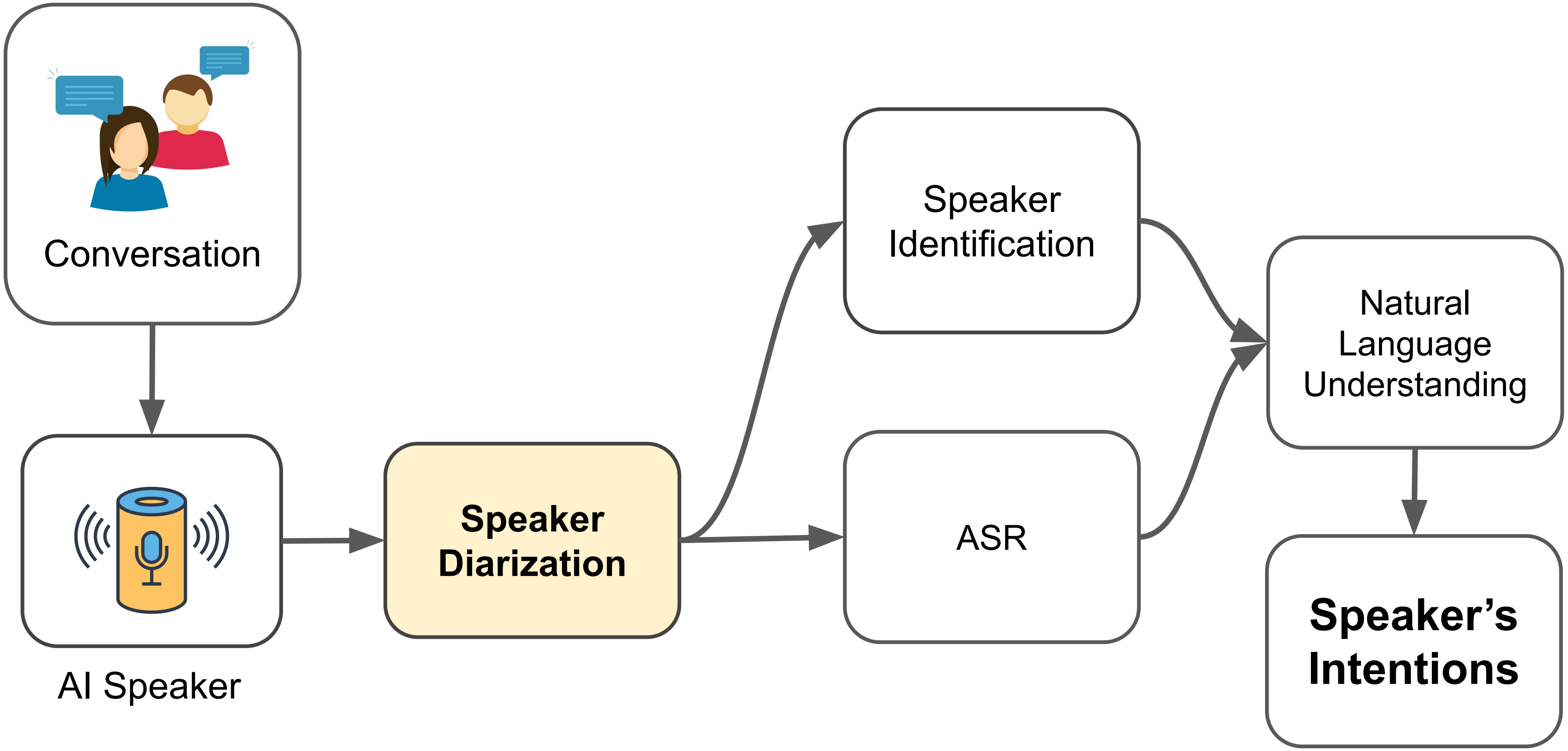
Chapter 3

Diarization Overview

Part-2

The State of Speaker Diarization

Diarization in Conversational AI



Diarization in Conversational AI

Speaker Diarization in Conversational AI

- Smart speakers In-home scenario (Amazon Alexa, Google Home etc)
- Targeting multi-human computer dialogues
- People have conversations themselves and the device listens to it
- Needs to keep track of who saying what

Andreas Stolcke (Amazon)

Emerging diarization technologies and services

Diarization in Conversational AI

Nvidia Jarvis

Alex : Show us a Japanese one.

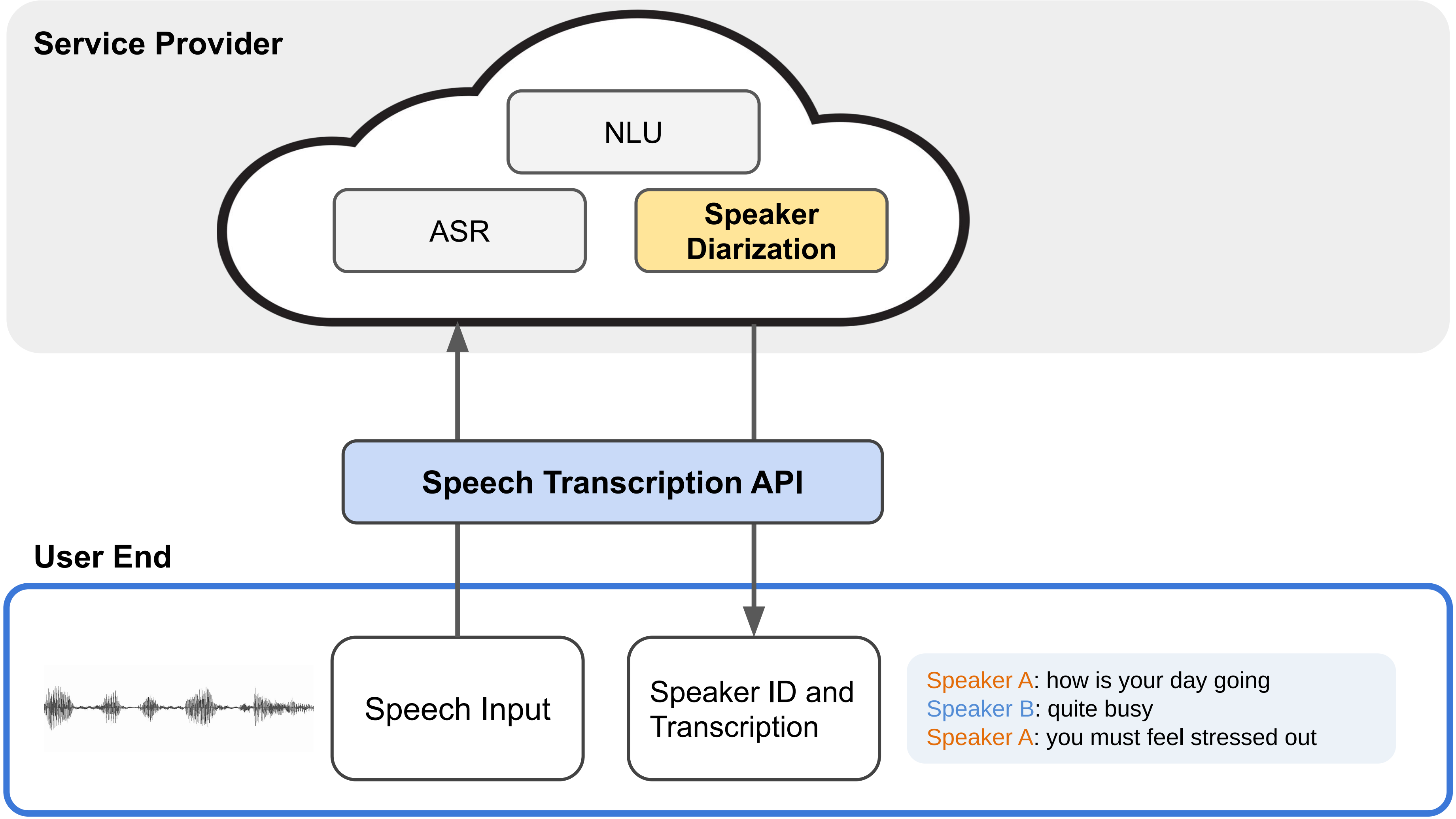
Jarvis : The nearest Japanese restaurant is Anikl's Sushi located at 3810 Mowry Avenue Fremont

Diarization as part of an e2e NLU pipeline - Diarization becomes a processing step

The Jarvis framework includes pretrained conversational AI models, tools, and optimized end-to-end services for speech, vision, and NLU tasks. In addition to AI services, Jarvis enables you to fuse vision, audio, and other sensor inputs simultaneously to deliver capabilities such as multi-user, multi-context conversations in applications such as virtual assistants, multi-user diarization, and call center assistants.

Emerging diarization technologies and services

Cloud based Transcription APIs



Speaker A: how is your day going
Speaker B: quite busy
Speaker A: you must feel stressed out

Cloud based Transcription APIs

Gakuto Kurata (IBM)

IBM’s Cloud based Transcription APIs

- IBM provides cloud based speech transcription API (Watson Speech to Text).

Specific applications are:

- Real time agent support system
- Automatic customer care service at contact center
- Speech analytics with natural language processing

Cloud based Transcription APIs

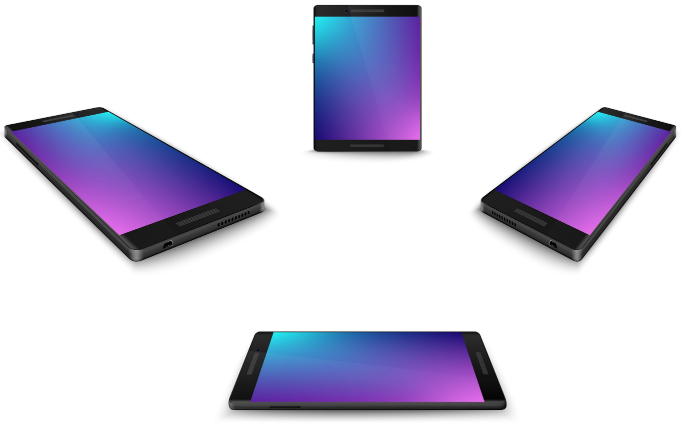
Rev.ai’s APIs are used in the following companies and applications:

- Media companies
- Meeting transcript
- Podcast transcript
- Public speaking training
- Interviews
- Market research
- Education (e.g. Zoom meetings)

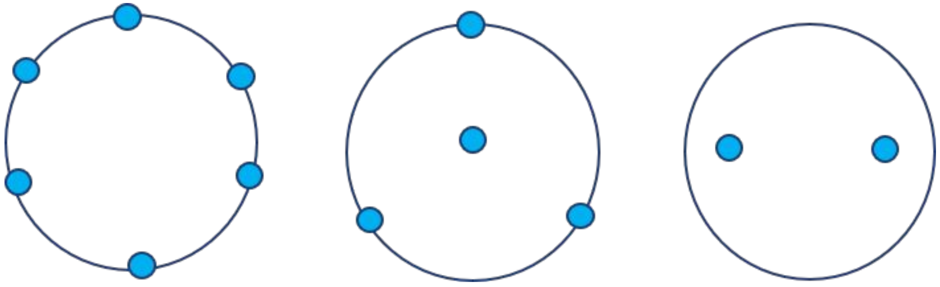
Miguel Jette (Rev.ai)

Diarization with Multi-Devices and Multi-Microphones

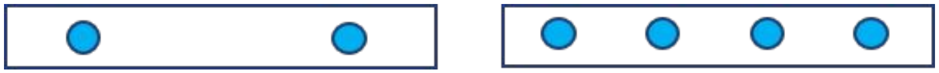
Synchronized Multi-device setup



Circular Arrays



Linear Arrays



- The advent of collaborative microphone network: Speaker Diarization and Multichannel ASR are done by synchronized multiple mobile device and take advantage of multiple signal sources.

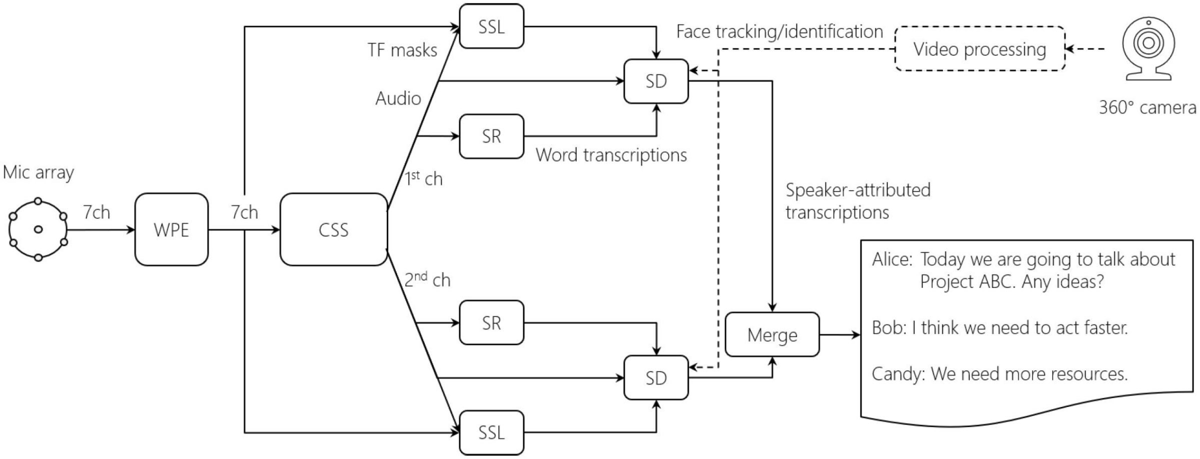
- Devices with multiple microphone setups (circular arrays and linear arrays) enable an enhanced speaker diarization performance and ASR accuracy.

Diarization with Multi-Devices and Multi-Microphones

Details of fixed geometry device

Separate-Recognize-Diarize Framework

- MIMO dereverberation is performed in real time
- Continuous Source Separation
- Speech recognition on separated signals
- Output words are input to Speaker Diarization module,
- Speaker labels are assigned, finally
- Speaker-annotated transcriptions from the N streams are merged



Alice: Today we are going to talk about Project ABC. Any ideas?

Bob: I think we need to act faster.

Candy: We need more resources.

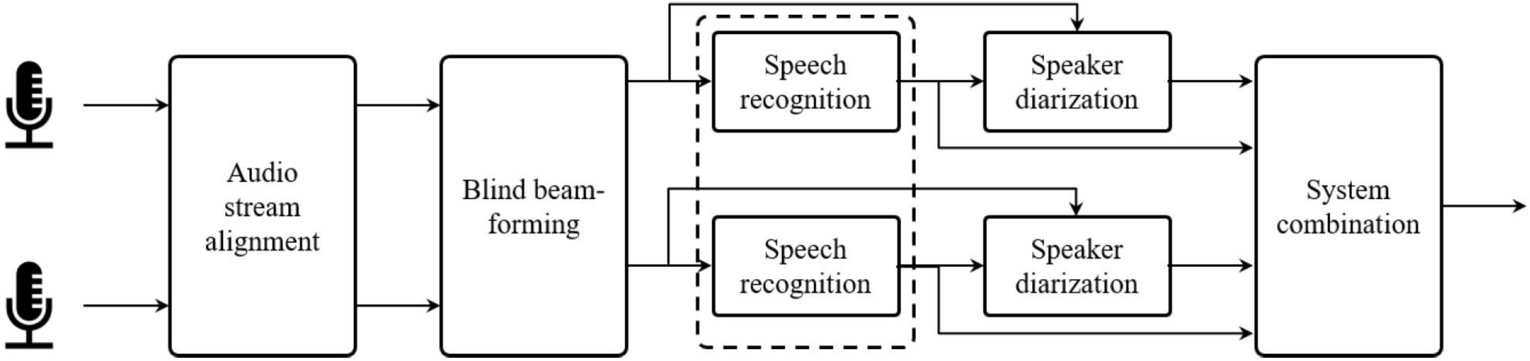
“Advances in Online Audio-Visual Meeting Transcription”, Yoshioka et al, arXiv:1912.04979, Dec. 2019

Diarization with Multi-Devices and Multi-Microphones

Ad-hoc Microphone Arrays

Processing steps:

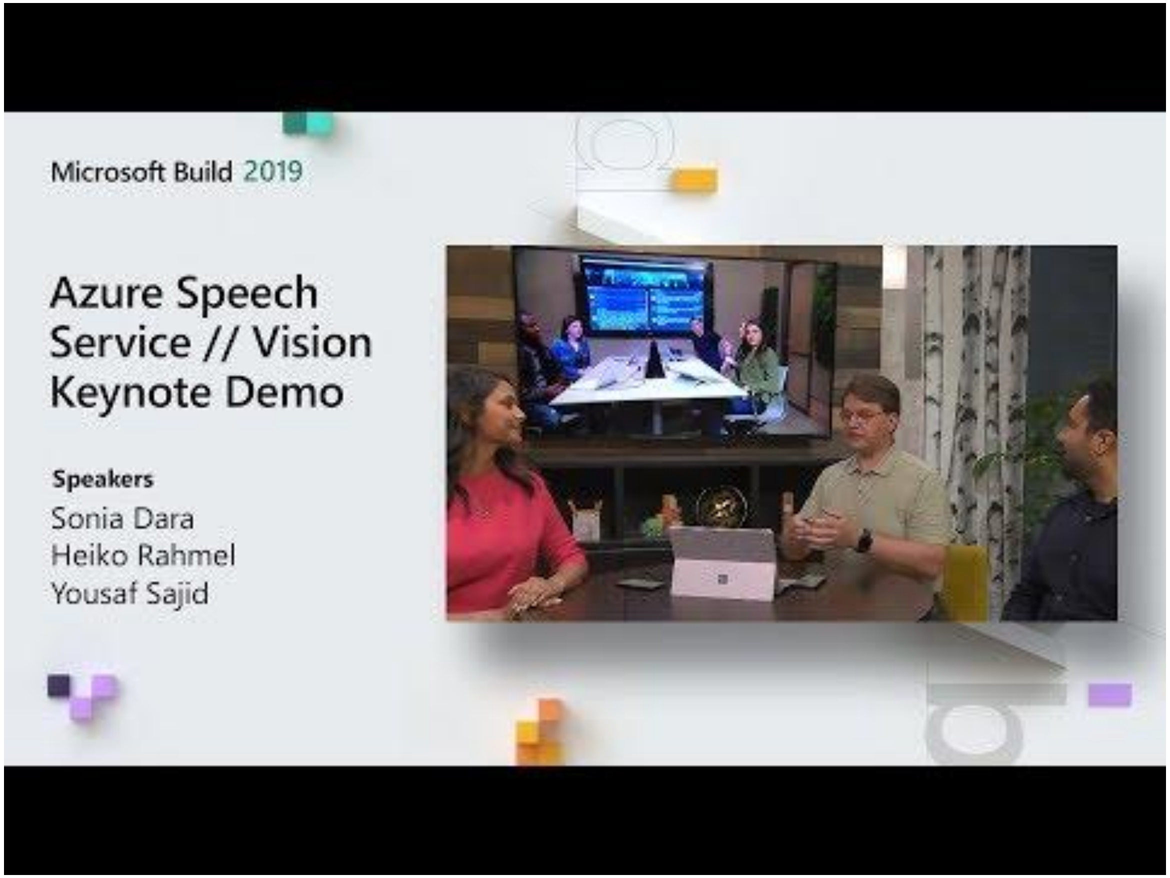
- Audio alignment
- Beamforming
- Speech recognition: Separate streams or multi-channel Acoustic Models
- Online svstem combination



“Meeting Transcription Using Virtual Microphone Arrays”, Yoshioka et al, arXiv:1905.02545, July 2019

Diarization with Multi-Devices and Multi-Microphones

Meeting Scenarios: Microsoft Azure Speech Service



Microsoft Build 2019

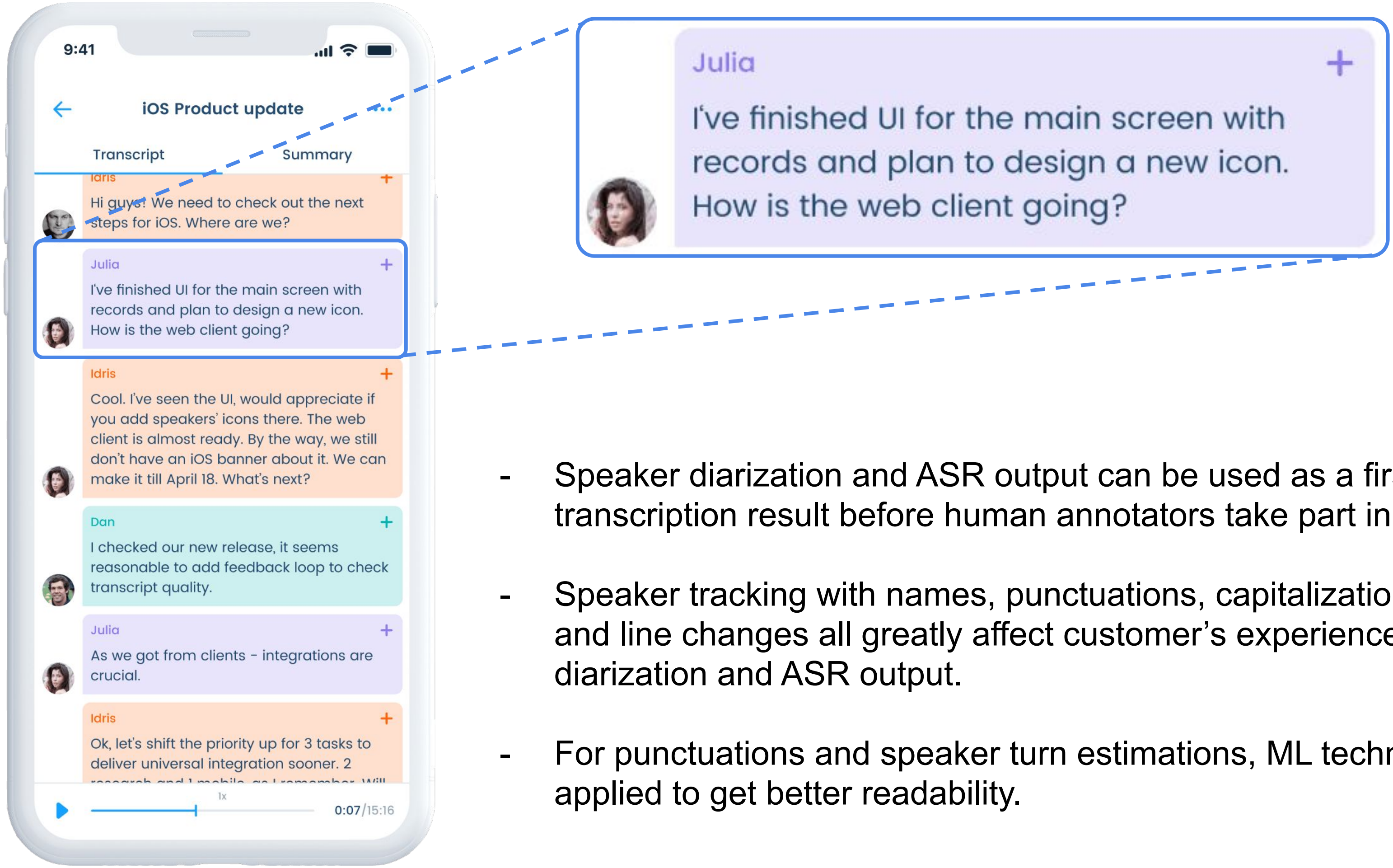
Azure Speech Service // Vision Keynote Demo

Speakers

Sonia Dara
Heiko Rahmel
Yousaf Sajid

Meeting transcriptions with ad-hoc microphone arrays

Diarization with Better Readability



9:41

iOS Product update

Transcript Summary

Idris
Hi guys! We need to check out the next steps for iOS. Where are we?

Julia
I've finished UI for the main screen with records and plan to design a new icon. How is the web client going?

Idris
Cool. I've seen the UI, would appreciate if you add speakers' icons there. The web client is almost ready. By the way, we still don't have an iOS banner about it. We can make it till April 18. What's next?

Dan
I checked our new release, it seems reasonable to add feedback loop to check transcript quality.

Julia
As we got from clients - integrations are crucial.

Idris
Ok, let's shift the priority up for 3 tasks to deliver universal integration sooner. 2 research and 1 mobile, as I remember. Will

1x 0:07/15:16

Julia
I've finished UI for the main screen with records and plan to design a new icon. How is the web client going?

- Speaker diarization and ASR output can be used as a first pass transcription result before human annotators take part in
- Speaker tracking with names, punctuations, capitalizations, spaces and line changes all greatly affect customer's experience on speaker diarization and ASR output.
- For punctuations and speaker turn estimations, ML techniques are applied to get better readability.

Diarization with Better Readability

- Speaker diarization is very important for customer satisfaction in speech transcript service.

- Better readability is crucial to to realize speech analytics and heavily affects customer satisfaction.

Gakuto Kurata (IBM)

Diarization with Better Readability

- “Revers” are transcribers at Rev.ai.
- Speaker diarization helps transcriptors to improve the final transcript result.
- If diarization goes wrong, it will make the transcription work very challenging.
- ASR accuracy and diarization accuracy are the two most important aspects for the final speech transcript result.

Miguel Jette (Rev.ai)

Next Generation diarization Applications

Domain specific Applications: Child speech

Healthcare domain

- We want to know the dynamics of spoken interaction
- How much a child talks to its caregivers (mon, dad or family members)?
- e.g. Autism spectrum disorder
- Separating child’s speech from caregiver’s speech and other background noise is the key part for this application.

Demo Video of Autism Spectrum Disorder

Shri Narayanan (USC)

Next Generation Diarization Applications

Diarization for media indexing: Gender bias study in movies

Demo Video of gender bias analysis

Shri Narayanan (USC)

Securities and intelligent robot

- Tracking and understanding multi-speaker activities for security concern
- Intelligent robot, understanding situations where multiple people interact in an informal manner

Shinji Watanabe (JHU)

Next Generation Diarization Applications

Speaker diarization for video games

- Entertainment is an emerging field of application of speaker diarization technology.
- There is a growing trend of online gaming and mobile gaming user base.
- Interactive multiplayer games require speaker diarization

Katrin Kirchhoff (Amazon)

Summary and Conclusions

Summary

Chapter 1: Diarization Overview

Chapter 2: Speaker Diarization and ASR

Part 1: Speaker diarization enhanced by ASR outputs
Part 2: Lexical information used in speaker diarization
Part 3: Joint modeling of speaker diarization and ASR

Chapter 3: Challenges and the State of Speaker Diarization

Part 1: Challenges in speaker diarization
Part 2: The State of speaker diarization

Conclusions

How far have we reached?

Speaker Diarization Systems

Human Listeners

- **Supervised tuning is required**
  - Segmentation, embedding and clustering
- **Only use single modality (audio)**
  - Acoustic features to embedding
- **No contextual information is involved**
  - Easily fails when audio feature degrades

- **Require less of explicit tuning**
  - Humans do not learn the task separately:
  - Humans act more like End-to-end system (Simultaneously optimized)
- **Exploit many different modalities**
  - Lexical context, role recognition etc.
- **Consider contextual information**
  - Very robust even if one modality degrades (ex. What if identical twins talk?)

End of the Presentation

Thank you!