Novel Architectures for Unsupervised Information Bottleneck Based Speaker Diarization of Meetings

Presenter: [†]Nauman Dawalatabad

Co-authors: [‡]Srikanth Madikeri, [†]C. Chandra Sekhar and [†]Hema A. Murthy

[†]IIT Madras, India
[‡]Idiap Research Institute, Switzerland

IEEE/ACM Transactions on Audio, Speech and Language Processing, v(29), 2021

Outline

- Diarization and its Applications
- Information Bottleneck (IB) based system
- Varying length segment initialization for IB based system (VarIB)
- Two-pass IB (TPIB) based system and VarTPIB system
- Results
- Conclusion

Diarization and Application

Speaker Diarization

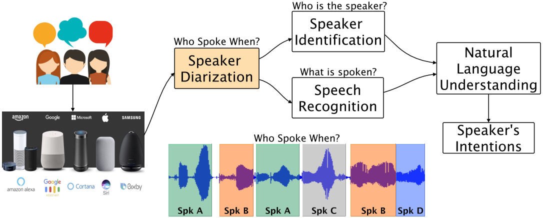
Given a conversation audio, a speaker diarization system answers the question of “Who Spoke When?”

Diarization and Application

Speaker Diarization

Given a conversation audio, a speaker diarization system answers the question of “Who Spoke When?”

Applications in Conversational AI



Who Spoke When?
Speaker Diarization

Who is the speaker?
Speaker Identification

What is spoken?
Speech Recognition

Natural Language Understanding

Speaker's Intentions

Who Spoke When?

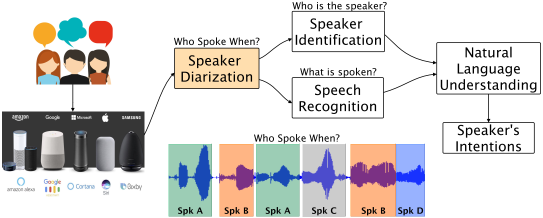
Conversational AI

Diarization and Application

Speaker Diarization

Given a conversation audio, a speaker diarization system answers the question of “Who Spoke When?”

Applications in Conversational AI



Who Spoke When?

Speaker Diarization

Who is the speaker?

Speaker Identification

What is spoken?

Speech Recognition

Natural Language Understanding

Speaker's Intentions

Who Spoke When?

Conversational AI

- Keyword spotting
- Source separation
- Peer-led team learning
- Professor life analysis
- Health care
- Marmoset vocalization

Challenges and Major Contributions

Major Challenges in Speaker Diarization

- Initialization of segments for clustering for bottom-up clustering.
- Obtaining speaker discriminative features.
- Deciding on the number of speakers.
- Detecting the overlapped speaker segments.

Challenges and Major Contributions

Major Challenges in Speaker Diarization

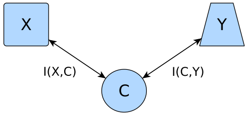
- Initialization of segments for clustering for bottom-up clustering.
- Obtaining speaker discriminative features.
- Deciding on the number of speakers.
- Detecting the overlapped speaker segments.

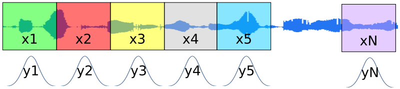
Major Contributions

1. Improve segment initialization of IB based approach.
2. Obtain a meeting specific speaker discriminative features using two-pass approach.

Information Bottleneck (IB) based speaker diarization

Unsupervised IB





Random variables X, Y, and C for Speech

- X represents segments in an audio – $\{x_1, x_2, \ldots, x_N\}$
- Y represents the Gaussian components – $\{y_1, y_2, \ldots, y_N\}$
- C represents the clusters made from X – $\{c_1, c_2, \ldots, c_m\}$, $m \leq N$

Maximize $\mathcal{F}$

$$\mathcal{F} = I(Y;C) - \frac{1}{\beta} I(C;X)$$

Key points

- Cluster segment posteriors $P(Y|X)$.
- Stopping NMI = $\frac{I(Y;C)}{I(X;Y)}$.

VarIB Approach

Motivation behind the proposed approach

- Current IB based system make use of uniform segmentation.
- Uniform segmentation may not be the best solution.
- Hence, proper segment initialization is needed.

VarIB Approach

Motivation behind the proposed approach

- Current IB based system make use of uniform segmentation.
- Uniform segmentation may not be the best solution.
- Hence, proper segment initialization is needed.
- The speaking rate can vary significantly across different speakers.
- Speaker information can be distributed uniformly across the segments.

VarIB Approach

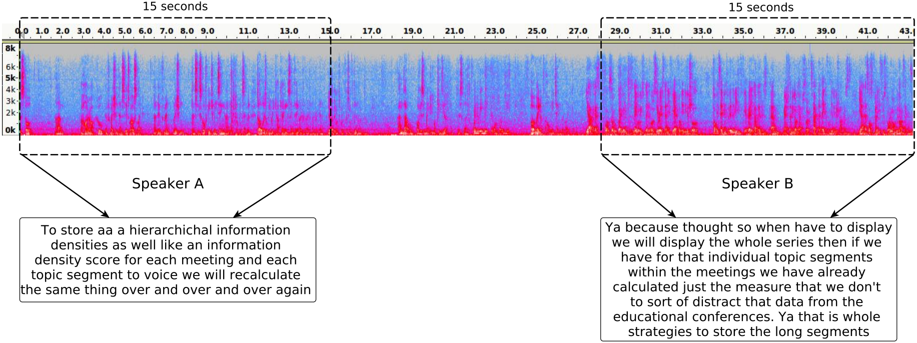
Motivation behind the proposed approach

- Current IB based system make use of uniform segmentation.
- Uniform segmentation may not be the best solution.
- Hence, proper segment initialization is needed.
- The speaking rate can vary significantly across different speakers.
- Speaker information can be distributed uniformly across the segments.

Objective

To distribute number of phonemes equally across the segments.

Different Speaking Rate



Speaker A

To store aa a hierarchichal information densities as well like an information density score for each meeting and each topic segment to voice we will recalculate the same thing over and over and over again

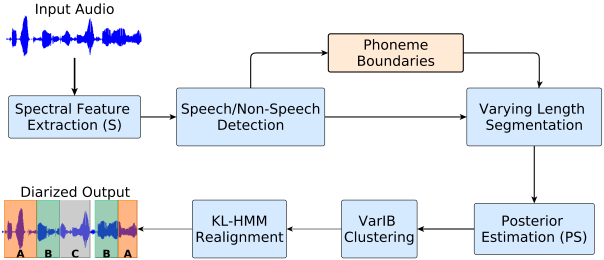
Speaker B

Ya because thought so when have to display we will display the whole series then if we have for that individual topic segments within the meetings we have already calculated just the measure that we don't to sort of distract that data from the educational conferences. Ya that is whole strategies to store the long segments

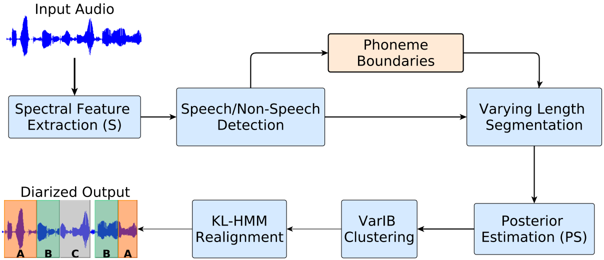
Varying Speaking Rate

- Varies across speakers.
- It can also varying within a speaker depending on his/her mood or current situation.

VarIB System


Input Audio
Spectral Feature Extraction (S)
Speech/Non-Speech Detection
Phoneme Boundaries
Varying Length Segmentation
Posterior Estimation (PS)
VarIB Clustering
KL-HMM Realignment
Diarized Output
A
B
C
B
A

VarIB System


Input Audio
Spectral Feature Extraction (S)
Speech/Non-Speech Detection
Phoneme Boundaries
Varying Length Segmentation
Posterior Estimation (PS)
VarIB Clustering
KL-HMM Realignment
Diarized Output
A
B
C
B
A

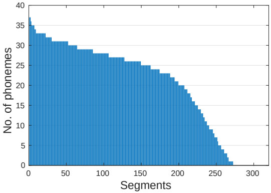
Optimization:

$$\mathcal{F}_v = I(Y;C) - \frac{1}{\beta} I(C;X_v)$$

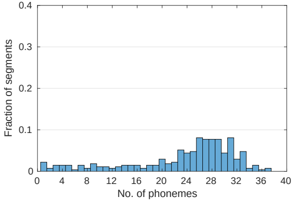
Posterior Estimation:

$$P(y_i|f_k) = \frac{a_i \mathcal{N}(f_k, \mu_i, \Sigma_i)}{\sum_{j=1}^{N} a_j \mathcal{N}(f_k, \mu_j, \Sigma_j)}$$

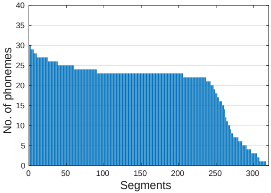
Distribution of Phonemes in VarIB Initialization



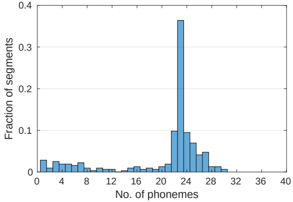
(a) No. of phonemes in fixed length segments



(b) Distribution for fixed length segments in IB



(c) No. of phonemes in varying length segments



(d) Distribution for varying length segments in VarIB

Two-pass IB (TPIB) based approach

Motivation behind the proposed system

- Current unsupervised systems do not make use of any discriminative feature information.

Two-pass IB (TPIB) based approach

Motivation behind the proposed system

- Current unsupervised systems do not make use of any discriminative feature information.
- One can make use of the discriminative information present in the output of the diarization system.

Two-pass IB (TPIB) based approach

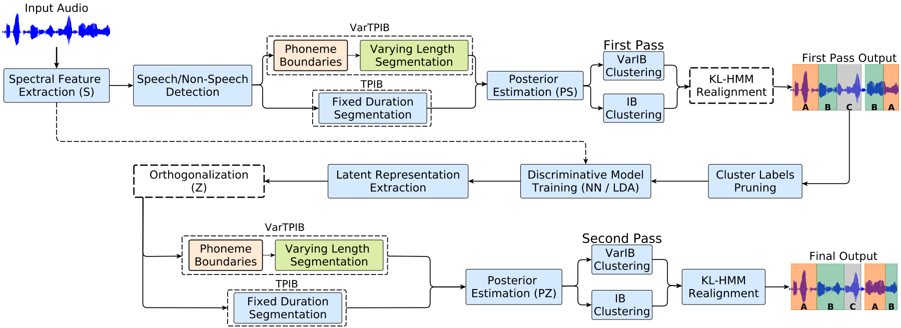
Motivation behind the proposed system

- Current unsupervised systems do not make use of any discriminative feature information.
- One can make use of the discriminative information present in the output of the diarization system.

Objective

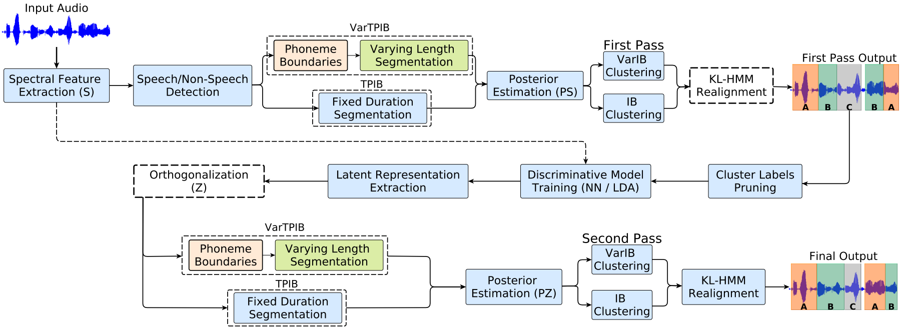
Introduce speaker discrimination model and keep the overall system unsupervised.

Two-pass IB (TPIB) based approach


Input Audio
Spectral Feature Extraction (S)
Speech/Non-Speech Detection
VarTPIB
Phoneme Boundaries
Varying Length Segmentation
TPIB
Fixed Duration Segmentation
Posterior Estimation (PS)
First Pass
VarIB Clustering
IB Clustering
KL-HMM Realignment
First Pass Output
A
B
C
B
A
Cluster Labels Pruning
Discriminative Model Training (NN / LDA)
Latent Representation Extraction
Orthogonalization (Z)
VarTPIB
Phoneme Boundaries
Varying Length Segmentation
TPIB
Fixed Duration Segmentation
Posterior Estimation (PZ)
Second Pass
VarIB Clustering
IB Clustering
KL-HMM Realignment
Final Output
A
B
C
A
B

Two-pass IB (TPIB/VarTPIB) based Speaker Diarization System.

Two-pass IB (TPIB) based approach


Input Audio
Spectral Feature Extraction (S)
Speech/Non-Speech Detection
VarTPIB
Phoneme Boundaries
Varying Length Segmentation
TPIB
Fixed Duration Segmentation
Posterior Estimation (PS)
First Pass
VarIB Clustering
IB Clustering
KL-HMM Realignment
First Pass Output
A
B
C
B
A
Cluster Labels Pruning
Discriminative Model Training (NN / LDA)
Latent Representation Extraction
Orthogonalization (Z)
VarTPIB
Phoneme Boundaries
Varying Length Segmentation
TPIB
Fixed Duration Segmentation
Posterior Estimation (PZ)
Second Pass
VarIB Clustering
IB Clustering
KL-HMM Realignment
Final Output
A
B
C
A
B

Two-pass IB (TPIB/VarTPIB) based Speaker Diarization System.

Key point

- Discriminative features extracted based on current recording.

Results on all Datasets

Diarization error rates for different systems.

System	Segment Initialization	Discriminative Model(s)	Features	Dev	Test Set			
				RT-04Dev	RT-04Eval	RT-05Eval	AMI-1	AMI-2
IB	Fixed	-	MFCC	15.1	13.5	16.4	17.9	23.5
Xvector+AHC+VB (Supervised, 5000 hours)	-	-	xvectors	10.4	10.9	10.4	9.7	10.5
			Proposed Systems					
VarIB	Varying	-	MFCC	12.3	12	15.3	17.8	22.6
TPIB	Fixed	MLFFNN	LF_{NN}	14.2	12.6	14.2	16.1	23.6
		LDA	LF_{LDA}	14.7	11.6	13.2	15.7	24.5
		MLFFNN+LDA	$LF_{NN} + LF_{LDA}$ (0.2,0.8)	13.1	12.6	12.6	15.4	21.9
		MLFFNN+LDA	$LF_{NN} + LF_{LDA}$ (Avg.)	14.2	12.4	14.5	16.3	22.2
VarTPIB	Varying	MLFFNN	LF_{NN}	12	**9.9**	14.2	17.5	**20.9**
		LDA	LF_{LDA}	13.8	12.8	**12.5**	14.8	21.3
		MLFFNN+LDA	$LF_{NN} + LF_{LDA}$ (0.6,0.4)	**11.6**	11.7	15.1	**13.2**	21.1
		MLFFNN+LDA	$LF_{NN} + LF_{LDA}$ (Avg.)	12.6	11.9	13.9	15.1	21.1

Results on all Datasets

Diarization error rates for different systems.

System	Segment Initialization	Discriminative Model(s)	Features	Dev	Test Set			
				RT-04Dev	RT-04Eval	RT-05Eval	AMI-1	AMI-2
IB	Fixed	-	MFCC	15.1	13.5	16.4	17.9	23.5
Xvector+AHC+VB (Supervised, 5000 hours)	-	-	xvectors	10.4	10.9	10.4	9.7	10.5
Proposed Systems								
VarIB	Varying	-	MFCC	12.3	12	15.3	17.8	22.6
TPIB	Fixed	MLFFNN	LF_{NN}	14.2	12.6	14.2	16.1	23.6
		LDA	LF_{LDA}	14.7	11.6	13.2	15.7	24.5
		MLFFNN+LDA	$LF_{NN} + LF_{LDA}$ (0.2,0.8)	13.1	12.6	12.6	15.4	21.9
		MLFFNN+LDA	$LF_{NN} + LF_{LDA}$ (Avg.)	14.2	12.4	14.5	16.3	22.2
VarTPIB	Varying	MLFFNN	LF_{NN}	12	**9.9**	14.2	17.5	**20.9**
		LDA	LF_{LDA}	13.8	12.8	**12.5**	14.8	21.3
		MLFFNN+LDA	$LF_{NN} + LF_{LDA}$ (0.6,0.4)	**11.6**	11.7	15.1	**13.2**	21.1
		MLFFNN+LDA	$LF_{NN} + LF_{LDA}$ (Avg.)	12.6	11.9	13.9	15.1	21.1

Runtimes in RTF.

System	RTF (x10)
IB	0.74
Xvector	2.13
VarIB	0.82
TPIB-NN	2.44
TPIB-LDA	1.42
VarTPIB-NN	2.58
VarTPIB-LDA	1.61

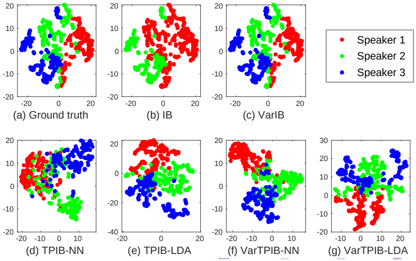
Results on all Datasets

Diarization error rates for different systems.

System	Segment Initialization	Discriminative Model(s)	Features	Dev	Test Set			
				RT-04Dev	RT-04Eval	RT-05Eval	AMI-1	AMI-2
IB	Fixed	-	MFCC	15.1	13.5	16.4	17.9	23.5
Xvector+AHC+VB (Supervised, 5000 hours)	-	-	xvectors	10.4	10.9	10.4	9.7	10.5
		Proposed Systems						
VarIB	Varying	-	MFCC	12.3	12	15.3	17.8	22.6
TPIB	Fixed	MLFFNN	LF_{NN}	14.2	12.6	14.2	16.1	23.6
		LDA	LF_{LDA}	14.7	11.6	13.2	15.7	24.5
		MLFFNN+LDA	$LF_{NN} + LF_{LDA}$ (0.2,0.8)	13.1	12.6	12.6	15.4	21.9
		MLFFNN+LDA	$LF_{NN} + LF_{LDA}$ (Avg.)	14.2	12.4	14.5	16.3	22.2
VarTPIB	Varying	MLFFNN	LF_{NN}	12	**9.9**	14.2	17.5	**20.9**
		LDA	LF_{LDA}	13.8	12.8	**12.5**	14.8	21.3
		MLFFNN+LDA	$LF_{NN} + LF_{LDA}$ (0.6,0.4)	**11.6**	11.7	15.1	**13.2**	21.1
		MLFFNN+LDA	$LF_{NN} + LF_{LDA}$ (Avg.)	12.6	11.9	13.9	15.1	21.1

Runtimes in RTF.

System	RTF (x10)
IB	0.74
Xvector	2.13
VarIB	0.82
TPIB-NN TPIB-LDA	2.44 1.42
VarTPIB-NN VarTPIB-LDA	2.58 1.61



(a) Ground truth (b) IB (c) VarIB

(d) TPIB-NN (e) TPIB-LDA (f) VarTPIB-NN (g) VarTPIB-LDA

Summary

Conclusions

- Better segment initialization results in better diarization output.
- Recording-specific discriminative features are incorporated.
- VarIB in tandem with TPIB further improves the performance.

Summary

Conclusions

- Better segment initialization results in better diarization output.
- Recording-specific discriminative features are incorporated.
- VarIB in tandem with TPIB further improves the performance.

Possible Extensions

- Other sound units like syllables or even word level segments.
- Other discriminative models in TPIB.

Summary

Conclusions

- Better segment initialization results in better diarization output.
- Recording-specific discriminative features are incorporated.
- VarIB in tandem with TPIB further improves the performance.

Possible Extensions

- Other sound units like syllables or even word level segments.
- Other discriminative models in TPIB.

More Information

N. Dawalatabad, S. Madikeri, C. C. Sekhar, H. A. Murthy, “Novel Architectures for Unsupervised Information Bottleneck based Speaker Diarization of Meetings”, IEEE/ACM Transactions of Audio, Speech and Language Processing, vol. 29, pp. 14–29, 2021.

Thank You!

Question(s), Comment(s) and/or Suggestion(s)?