INCREMENTAL TRANSFER LEARNING IN TWO-PASS INFORMATION BOTTLENECK BASED SPEAKER DIARIZATION SYSTEM FOR MEETINGS

Nauman Dawalatabad[1], Srikanth Madikeri[2], C Chandra Sekhar[1], Hema A Murthy[1]

[1]Indian Institute of Technology Madras, India
[2]Idiap Research Institute, Martigny, Switzerland

Introduction

- Speaker Diarization: *"Who Spoke When?"*
- Unsupervised speaker diarization.
- Uses "Remember–Learn–Transfer" principle to transfer the learned information.
- Reduces the real time factor of two-pass system.

IB based Diarization

- A set of segments $\mathcal{X}$ in an audio is clustered into set of clusters $\mathcal{C}$ preserving the relevant information $\mathcal{Y}$. The objective function is given by

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

I is mutual information between the variables and β is a Lagrange multiplier.

Two-pass IB based diarization

- **First pass:** IB based diarization is performed to obtain relative speaker labels.
- **ANN Training & LSF Extraction:** ANN initialized with *random weights* is trained from scratch on the output boundary labels and the spectral features to obtain latent features (LSF).
- **Second pass:** The LSFs are used along with the spectral features in the second pass of IB system.

TPIB-ITL

- **Training seedANN:** A seedANN is trained from the first audio to be diarized by the system.
- **First pass:** IB based diarization is performed to obtain relative speaker labels.
- **Remember–Learn–Transfer:** ANN initialized with *weights from seedANN* is fine-tuned on the output boundary labels and the spectral features of the current recording to obtain LSF. Store the fine-tuned ANN for next recording.
- **Second pass:** A second pass of IB based clustering is performed.

Incremental Transfer Learning in TPIB

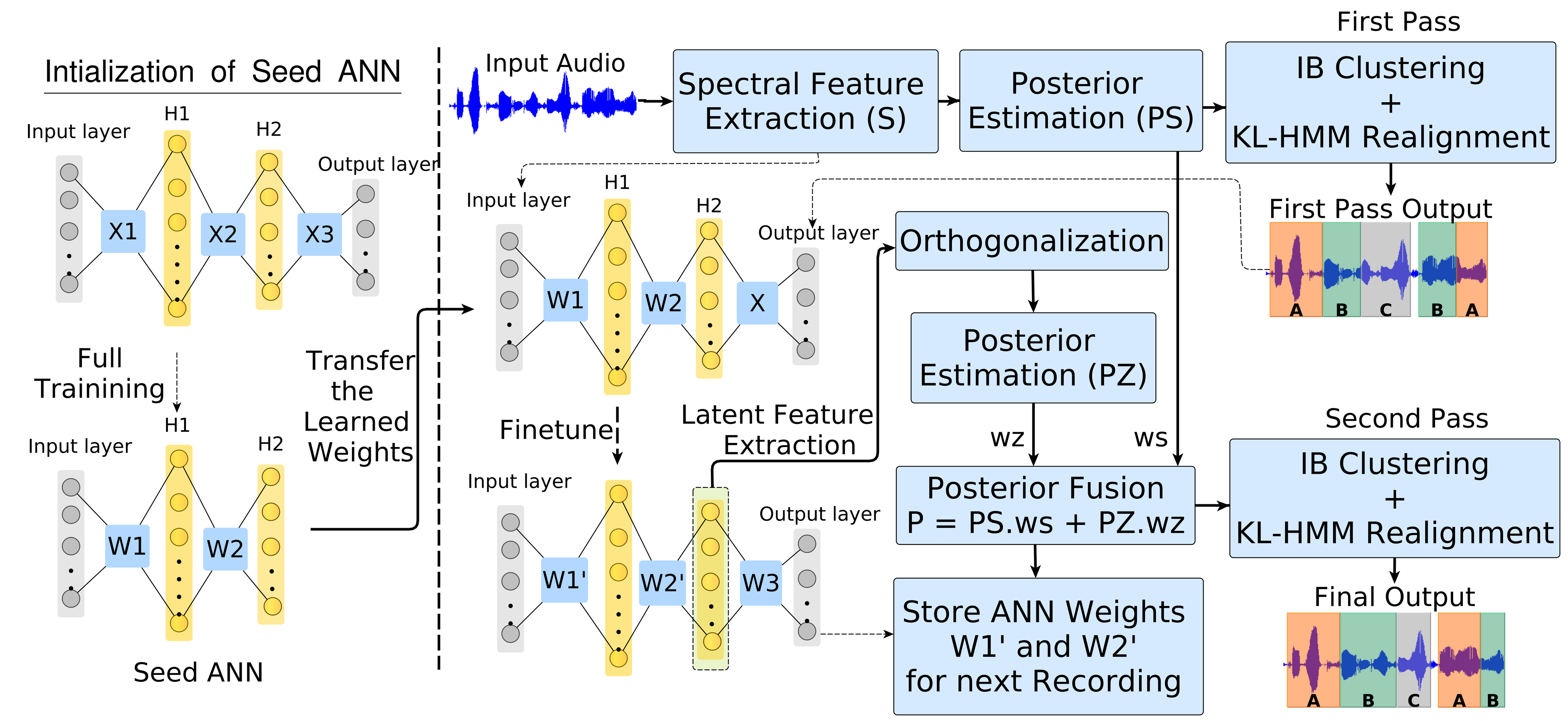


Figure : TPIB-ITL: Incremental Transfer Learning in TPIB

Important Result

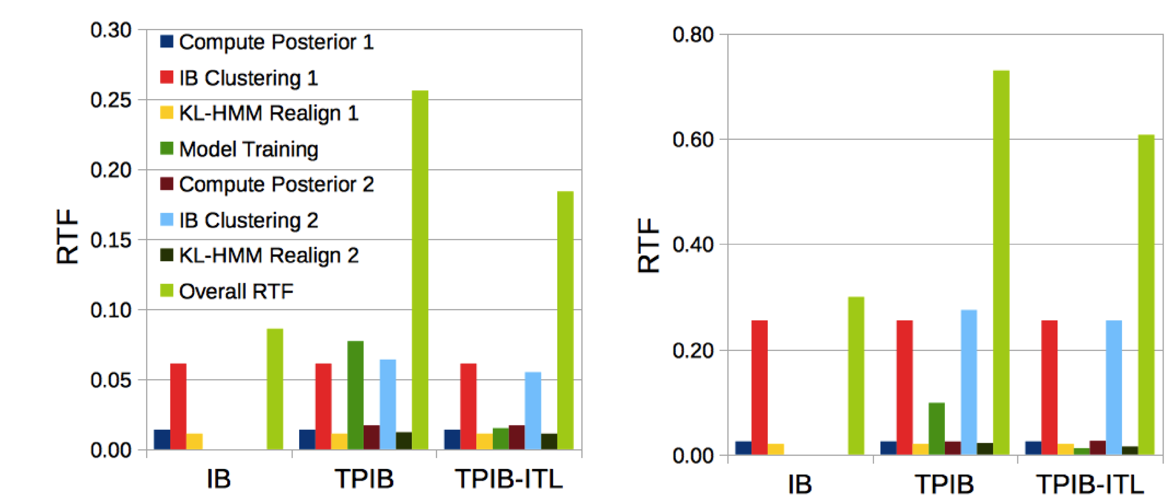
- TPIB-ITL uses "Remember–Learn–Transfer" principle to diarize new recordings.
- Retaining previous knowledge helps to reduce real time factor compared to TPIB system.

Speaker Error Rates

Table : Speaker Error Rate (SER) on different systems are mentioned. The feature fusing weights are mentioned in parentheses. Avg. denotes the average SER over all fusing weights combination. Best SER on both systems for each dataset is indicated in bold font.

System	Feature(s)	Dev. Set	Test Sets			
		RT-04Dev	RT-04Eval	RT-05Eval	AMI-1	AMI-2
IB	MFCC	15.1	13.5	16.4	17.9	23.5
TPIB	LSF	15.1	11.6	14.2	17.5	21.3
	MFCC+LSF (0.8, 0.2)	13.1	12.5	16.6	16.4	22.7
	MFCC+LSF (Avg.)	14.9	12.6	15.3	17.8	22.4
Proposed System						
TPIB-ITL	LSF	15.5	12.5	15.1	**17.5**	22
	MFCC+LSF (0.1, 0.9)	15.2	**12.2**	**15**	18	**21.2**
	MFCC+LSF (Avg.)	15.8	12.5	15.4	17.8	21.9
TPIB-ITL (Dev.)	LSF	15.5	12.9	**14.8**	17.5	22.1
	MFCC+LSF (0.1, 0.9)	15.2	**12.5**	15	**17.5**	**22**
	MFCC+LSF (Avg.)	15.8	13.3	15.6	17.8	22.5

Real Time Factors



(a) RT-05Eval Dataset (b) AMI-2 Dataset

Figure : RTFs of individual modules for all systems.

Overall RTF

Table : The RTF on different systems for different datasets are mentioned. Impr. denotes the relative improvement in RTF with respect to TPIB system.

Sys/Dataset	RT-04Dev	RT-04Eval	RT-05Eval	AMI-1	AMI-2
IB	0.070	0.081	0.086	0.241	0.304
TPIB	0.248	0.257	0.254	0.642	0.740
TPIB-ITL	0.175	0.172	0.180	0.485	0.605
Impr. (%)	29.44	**33.07**	29.13	24.45	18.24

Conclusion

- No separate training data is used.
- Recording-specific discrimination is achieved.
- Sequence of recordings does not affect the performance.
- TPIB-ITL also works when only development data in used in incremental learning phase.

References

[1] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard. An Information Theoretic Approach to Speaker Diarization of Meeting Data. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7):1382–1393, 2009.

[2] Nauman Dawalatabad, Srikanth Madikeri, C. Chandra Sekhar, and Hema A. Murthy. Two-Pass IB Based Speaker Diarization System Using Meeting-Specific ANN Based Features. In *Proceedings of INTERSPEECH*, pages 2199–2203, 2016.