Speech Understanding - Assignment 1

Cocktail Party Problem

Presented By :
Ravi Saxena(P23CS006)
Aditi Baheti(M23CSA001)

Under guidance of:
Prof Richa Singh

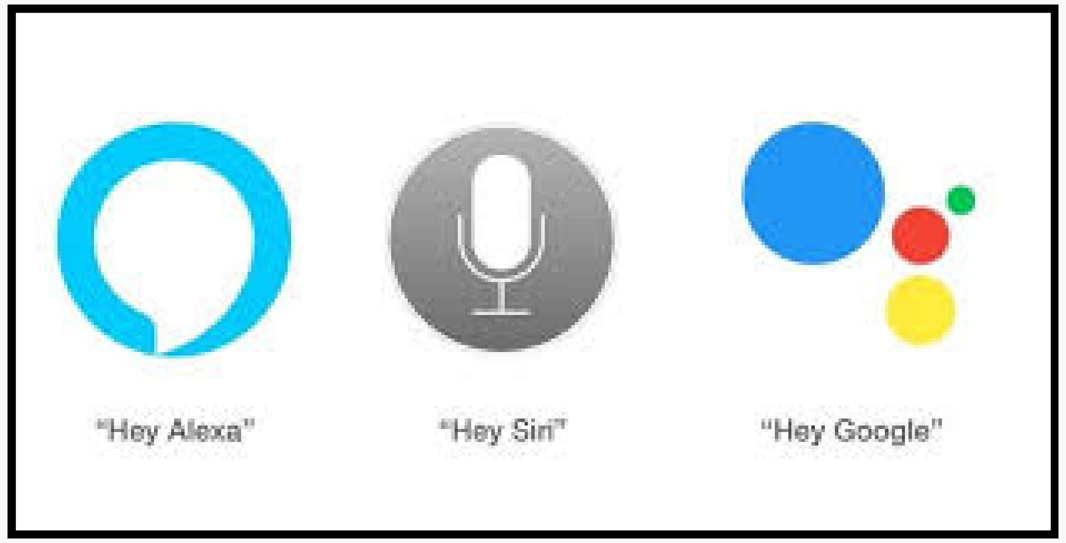
3rd February, 2025

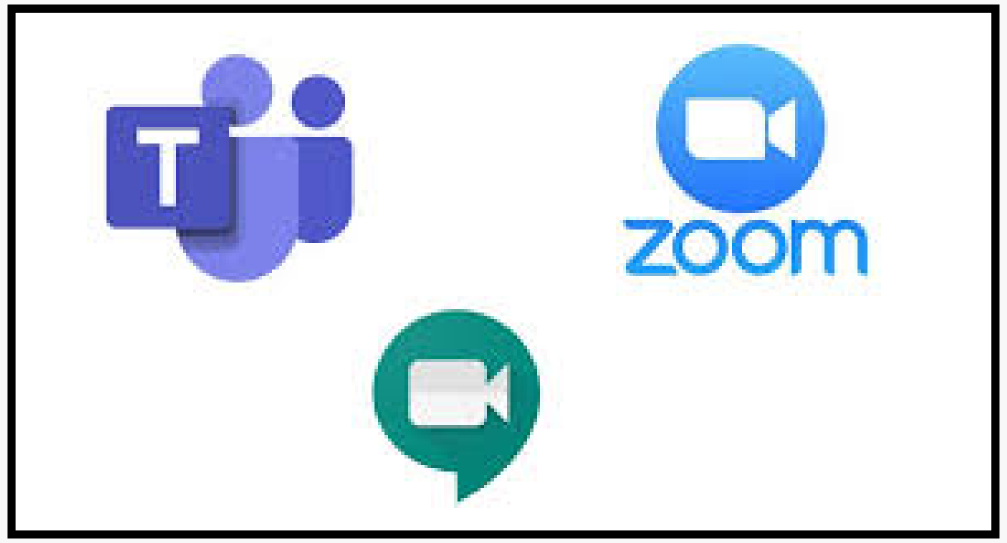
The Task and Its Real-World Importance

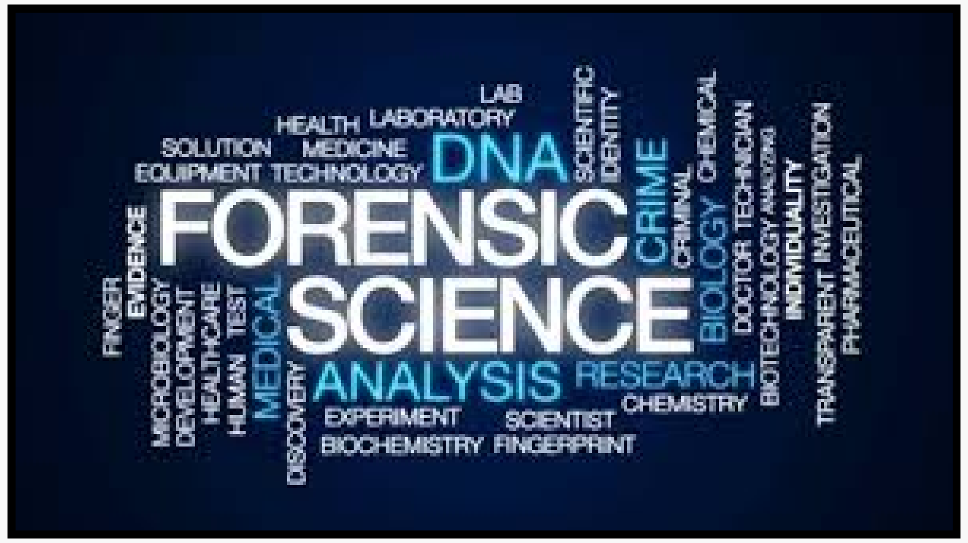
“Hearing the Right Voice & Forgetting the Wrong One”

Imagine being at a crowded party. Your brain naturally tunes into one voice. But what if AI could do the same—separate voices with precision?

Now, what if we take it one step further? What if AI could 'unlearn' a voice—removing it like it was never there, without affecting the rest?







SOTA : Separate and Reconstruct: Asymmetric Encoder-Decoder for Speech Separation

System	Params. (M)	MACs (G/s)	WSJ0-2Mix SI-SNRi (dB)	WSJ0-2Mix SDRi (dB)	WHAM! SI-SNRi (dB)	WHAM! SDRi (dB)	Libri2Mix SI-SNRi (dB)	Libri2Mix SDRi (dB)
Conv-TasNet [47]	5.1	10.5	15.3	15.6	12.7	-	12.2	12.7
SuDoRM-RF [70]	6.4	10.1	18.9	-	13.7	14.1	14.0	14.4
TDANet [42]	2.3	9.1	18.5	18.7	15.2	15.4	17.4	17.9
Sandglasset [38]	2.3	28.8	20.8	21.0	-	-	-	-
S4M [7]	3.6	38.4	20.5	20.7	-	-	16.9	17.4
SepReformer-T	3.5	10.4	22.4	22.6	17.2	17.5	19.7	20.2
SepReformer-S	4.3	21.3	23.0	23.1	17.3	17.7	20.6	21.0
DPRNN [45]	2.6	88.5	18.8	19.0	13.7	14.1	16.1	16.6
DPTNet [9]	2.7	102.5	20.2	20.3	14.9	15.3	16.7	17.1
Sepformer [66]	26.0	86.9	20.4	20.5	14.7	16.8	16.5	17.0
WaveSplit† [89]	29.0	-	21.0	21.2	16.0	16.5	16.6	17.2
A-FRCNN [32]	6.1	125.0	18.3	18.6	14.5	14.8	16.7	17.2
SFSRNet [60]	59.0	124.2	22.0	22.1	-	-	-	-
ISCIT† [51]	58.4	252.2	22.4	22.5	16.4	16.8	-	-
QDPN [59]	200.0	-	22.1	-	-	-	-	-
TF-GridNet [79]	14.5	460.8	23.5	23.6	-	-	-	-
SepReformer-B	14.2	39.8	23.8	23.9	17.6	18.0	21.6	21.9
SepReformer-M	17.3	81.3	24.2	24.4	17.8	18.1	22.0	22.2

1. **WSJ0-2Mix**: Contains 30 hours (train), 10 hours (validation), and 5 hours (evaluation).Mixtures are generated by randomly selecting different speakers and mixing them at random SNRs between -5 dB and 5 dB.
2. **WHAM!** : Noisy and noisy-reverberant versions of WSJ0-2Mix.It mixes speech with recorded noise from real-world environments.
3. **Libri2Mix**: Derived from LibriSpeech train-100 dataset.Mixtures are created with randomly selected target speech.

- Asymmetric Encoder-Decoder with Early Split
- Global-Local Transformer for Long range and Local Dependencies

Limitation : Max two speaker mixtures

https://arxiv.org/abs/2406.05983

SepTDA

Models	Domain	Path	#params (M)	ΔSI-SDR (dB)	ΔSDR (dB)
DPRNN [11]	Time	Dual	2.6	18.8	19.0
Gated DPRNN [24]	Time	Dual	7.5	20.1	20.4
DPTNet [12]	Time	Dual	2.7	20.2	20.6
SepFormer [14]	Time	Dual	26.0	20.4	20.5
Wavesplit [13]	Time	Single	29.0	21.0	21.2
QDPN [15]	Time	Q-Dual	200.0	22.1	-
$\text{SepEDA}_2{}^*$ [28]	Time	Triple	12.5	21.2	21.4
MossFormer(L)* [18]	Time	Single	42.1	22.8	-
TF-GridNet [17]	TF	Dual	14.5	23.5	23.6
SepTDA_2	Time	Triple	21.2	**23.7**	**23.5**
with $L = 12$	Time	Triple	21.2	**24.0**	**23.9**

Dataset : WSJ0-2Mix: Contains 30 hours (train), 10 hours (validation), and 5 hours (evaluation).Mixtures are generated by randomly selecting different speakers and mixing them at random SNRs between -5 dB and 5 dB

ConvTasNet

COMPARISON WITH OTHER METHODS ON WSJ0-2MIX DATASET

Method	Model size	Causal	SI-SNRi (dB)	SDRi (dB)
DPCL++ [5]	13.6M	×	10.8	–
uPIT-BLSTM-ST [7]	92.7M	×	–	10.0
DANet [8]	9.1M	×	10.5	–
ADANet [9]	9.1M	×	10.4	10.8
cuPIT-Grid-RD [50]	47.2M	×	–	10.2
CBLDNN-GAT [12]	39.5M	×	–	11.0
Chimera++ [10]	32.9M	×	11.5	12.0
WA-MISI-5 [11]	32.9M	×	12.6	13.1
BLSTM-TasNet [26]	23.6M	×	13.2	13.6
Conv-TasNet-gLN	**5.1M**	×	**15.3**	**15.6**
uPIT-LSTM [7]	46.3M	✓	–	7.0
LSTM-TasNet [26]	32.0M	✓	**10.8**	**11.2**
Conv-TasNet-cLN	**5.1M**	✓	10.6	11.0
IRM	–	–	12.2	12.6
IBM	–	–	13.0	13.5
WFM	–	–	13.4	13.8

Paper Link : ConvTasNet

Paper Link : SepTDA

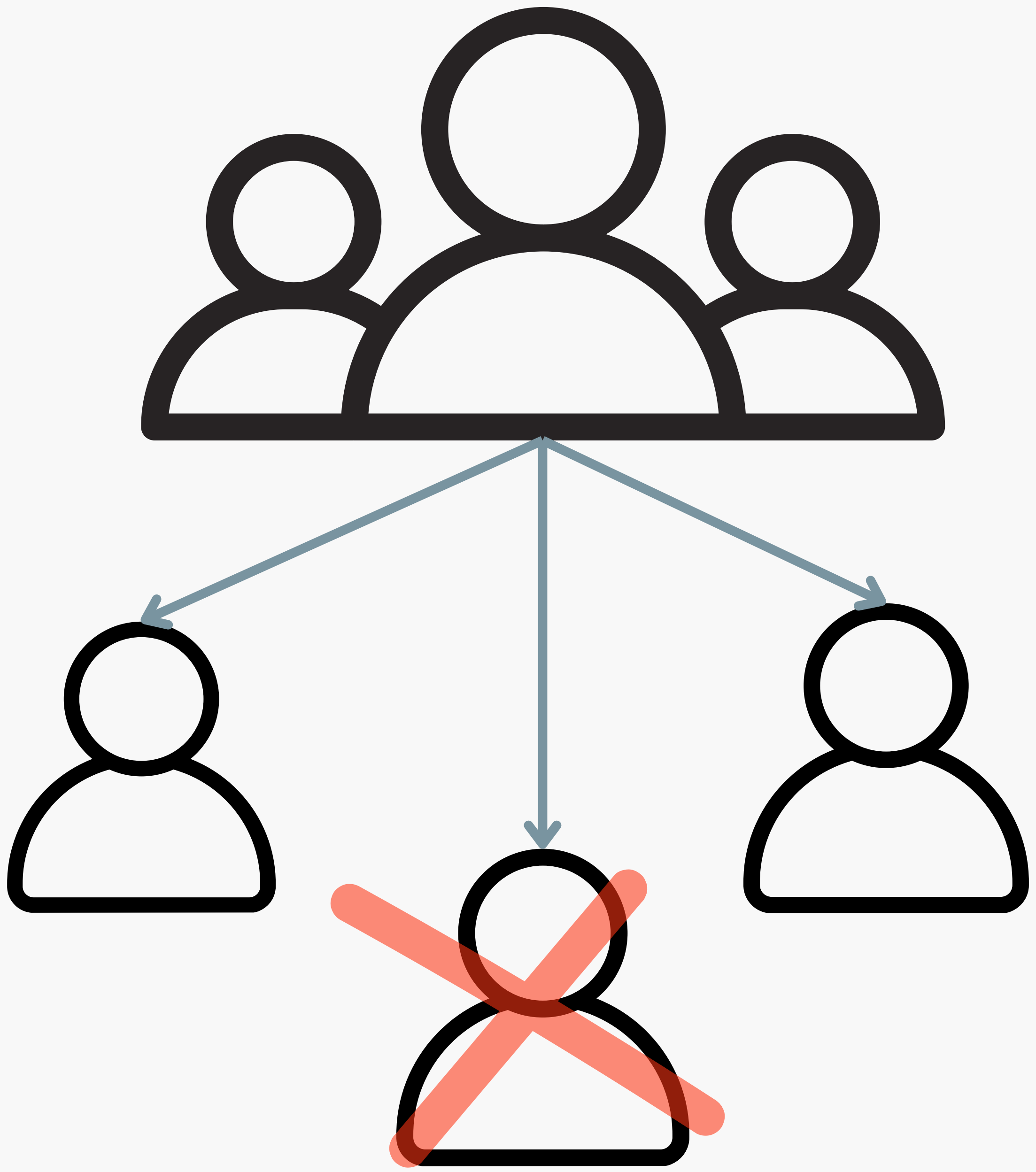
Strengths & Limitations of Existing Approaches

Approach 1:Cocktail Party Problem

- Separates mixed voices Doesn't "forget" voices—identities remainoduct sales going down because of some reason.

Approach 2: Machine Unlearning

- Removes specific voices from a model (AmnesiacML)
- Not designed for audio scenarios



Unique Solution: Combining Both

- Separate voices **(Cocktail Party)**
- Selectively erase voices **(Machine Unlearning)**
- Preserve speech clarity **(Accent-Aware Learning)**

[Link to Amnesiac ML]()

Open Challeneges and Research Opportunities

Lack of Selective Voice Forgetting

Accent & Speaker Bias in Speech AI

Trade-off Between Unlearning & Speech Quality

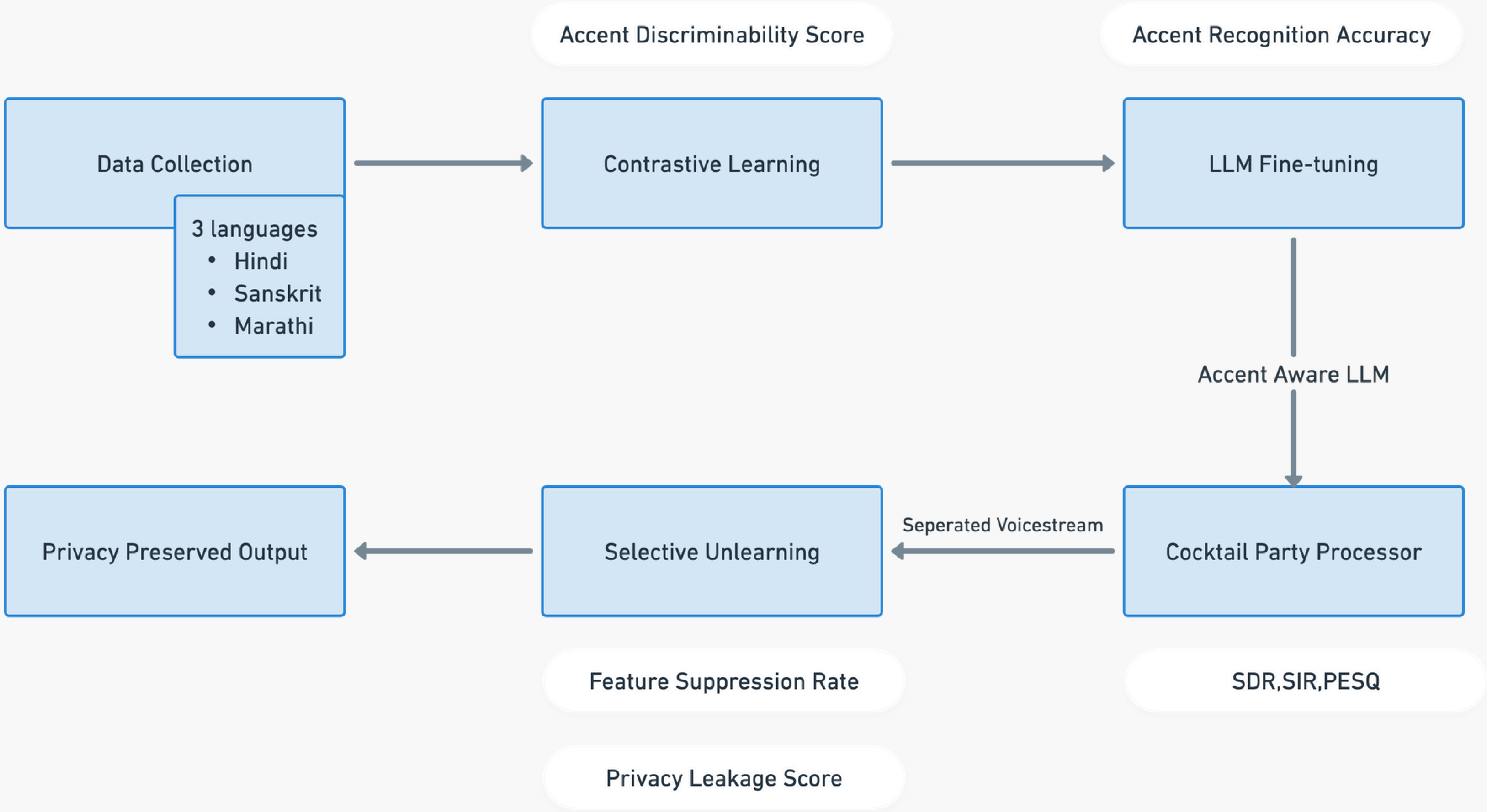
Integrated Voice Separation & Machine Unlearning

- Trying to combine Conv-TasNet for speech separation with AmnesiacML for targeted voice removal.
- Selectively removes speaker characteristics while preserving overall speech intelligibility.

Unified Loss Function for Separation & Unlearning

- Planning to introduce a combined loss function that jointly optimizes Conv-TasNet for separation and feature suppression for unlearning.
- Ensures a balanced trade-off between speech clarity and privacy preservation.

Approach


Accent Discriminability Score
Accent Recognition Accuracy
Data Collection
3 languages
• Hindi
• Sanskrit
• Marathi
Contrastive Learning
LLM Fine-tuning
Accent Aware LLM
Privacy Preserved Output
Selective Unlearning
Seperated Voicestream
Cocktail Party Processor
Feature Suppression Rate
SDR,SIR,PESQ
Privacy Leakage Score

References

1. Luo, Y. and Mesgarani, N., 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM transactions on audio, speech, and language processing, 27(8), pp.1256-1266.
2. Graves, L., Nagisetty, V. and Ganesh, V., 2021, May. Amnesiac machine learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 13, pp. 11516-11524).
3. Fazel-Zarandi, M. and Hsu, W.N., 2023, June. Cocktail Hubert: Generalized Self-Supervised Pre-Training for Mixture and Single-Source Speech. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.
4. Liu, X., Kong, Q., Zhao, Y., Liu, H., Yuan, Y., Liu, Y., Xia, R., Wang, Y., Plumbley, M.D. and Wang, W., 2024. Separate anything you describe. IEEE/ACM Transactions on Audio, Speech, and Language Processing.
5. Sai, S., Mittal, U., Chamola, V., Huang, K., Spinelli, I., Scardapane, S., Tan, Z. and Hussain, A., 2024. Machine unlearning: an overview of techniques, applications, and future directions. Cognitive Computation, 16(2), pp.482-506.

Thank you