Google

ASRU 2021 Cartagena | December 13th to 17th
IEEE AUTOMATIC SPEECH RECOGNITION AND UNDERSTANDING WORKSHOP

Multi-user VoiceFilter-Lite via Attentive Speaker Embedding

Rajeev V. Rikhye*, Quan Wang*, Qiao Liang, Yanzhang He, Ian McGraw

Abstract

Problem:

- Most *speaker conditioned speech models* only allow a single enrolled speaker

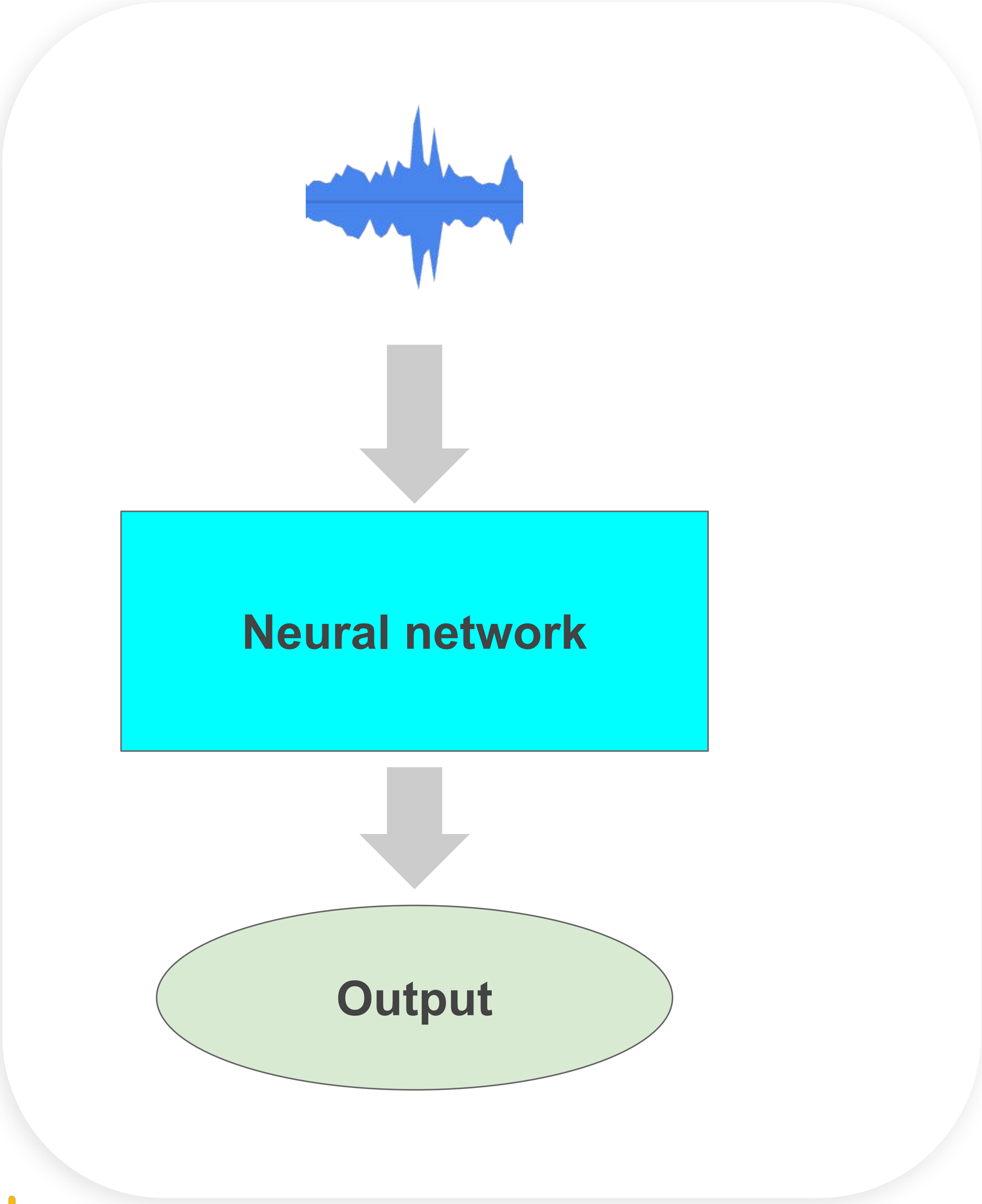
Our solution:

- A novel **attention mechanism** to identify which of the *N* enrolled users is speaking in a particular frame
- This **attentive embedding** can then be used with any speaker conditioned model like VoiceFilter-Lite, Personal Voice Activity Detection, or Personalized ASR

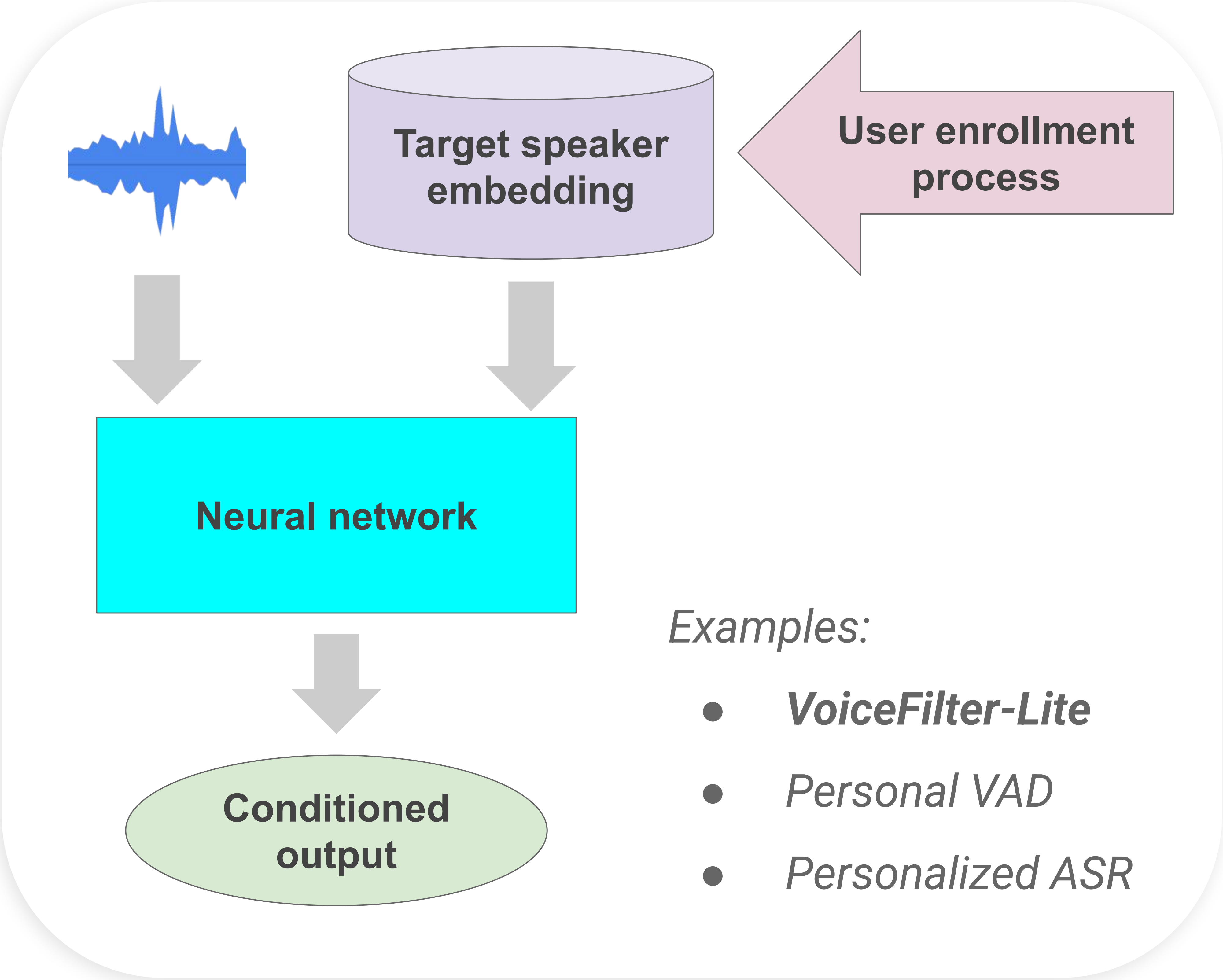
Experiments:

- **Multi-user VoiceFilter-Lite** significantly reduces speech recognition and speaker verification errors when there is overlapping speech, without affecting performance under other acoustic conditions

Speaker conditioned speech models



Generic speech model

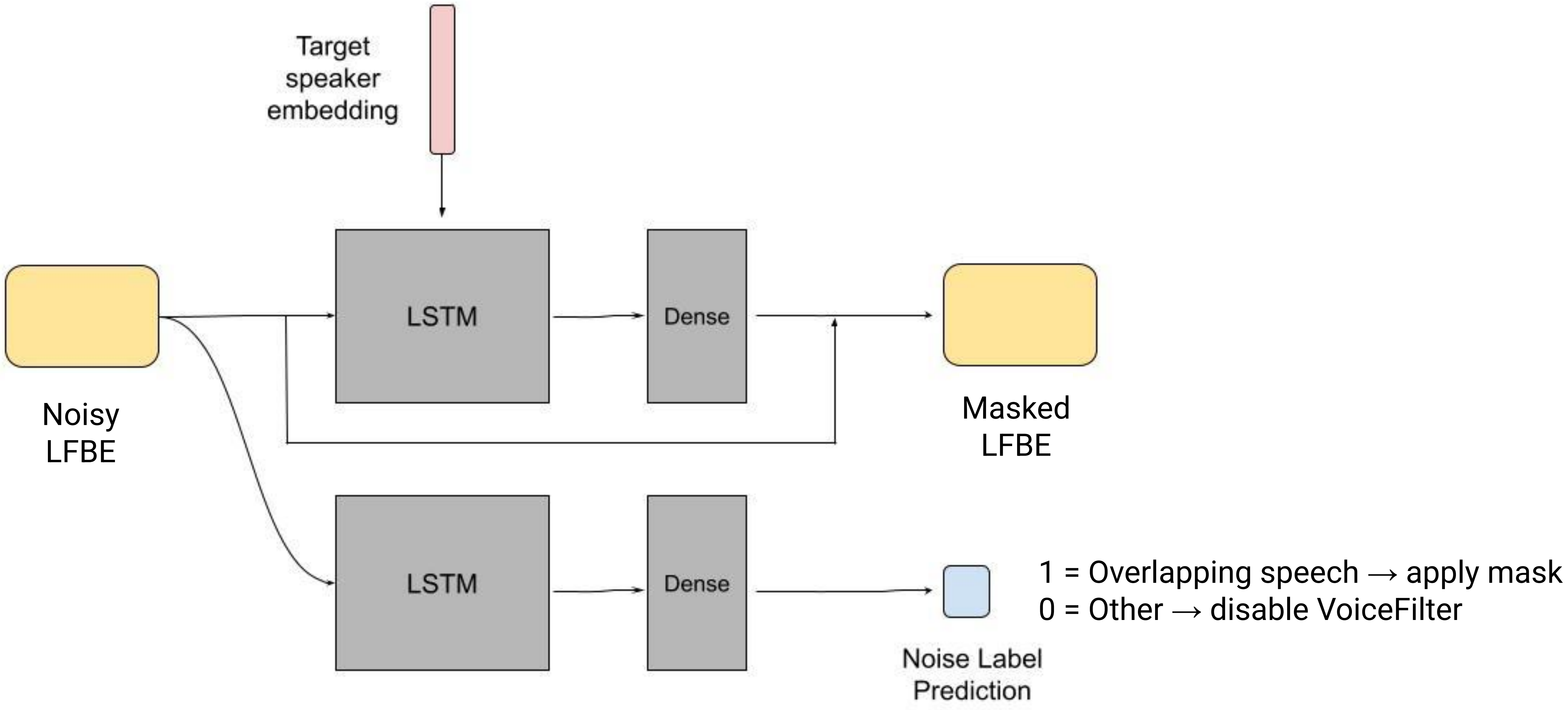


Speaker conditioned speech model

Multi-user VoiceFilter-Lite Model

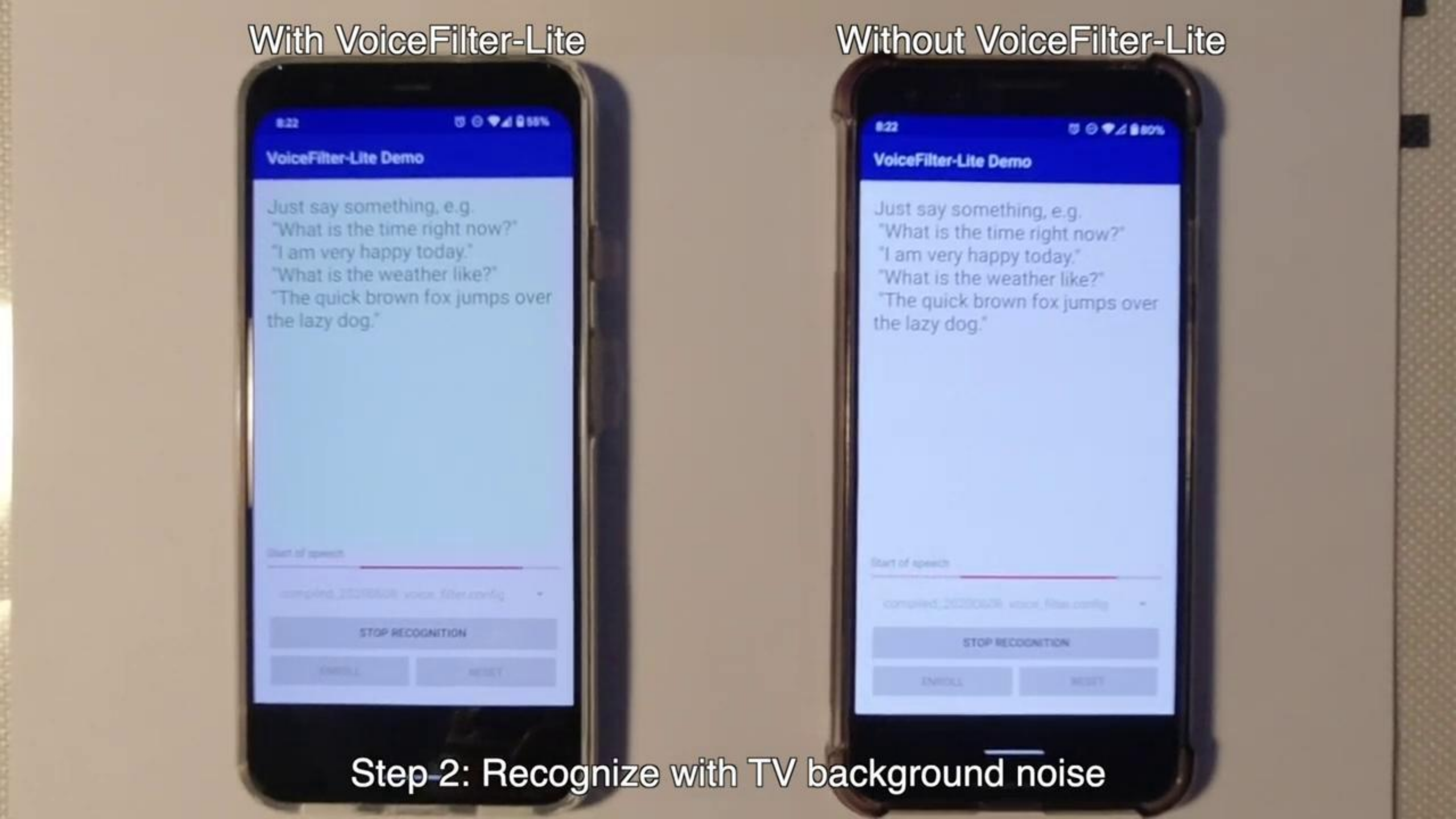
VoiceFilter-Lite enhances **target user** speech in multitalker environments

Model size: 2.7 MB



- The VoiceFilter-Lite (SUVF) [1] takes as input the target speaker embedding and a stacked log Mel filterbank energies (LFBE) and returns an “enhanced” LFBE and a noise label prediction.
- SUVF **suppresses** overlapping speech from non-enrolled users.

[1] Q. Wang, et al., “VoiceFilter-Lite: Streaming targeted voice separation for on-device speech recognition,” in Proc. Interspeech, 2020, pp. 2677–2681


With VoiceFilter-Lite
Without VoiceFilter-Lite
VoiceFilter-Lite Demo
Just say something, e.g.
"What is the time right now?"
"I am very happy today."
"What is the weather like?"
"The quick brown fox jumps over the lazy dog."
STOP RECOGNITION
VoiceFilter-Lite Demo
Just say something, e.g.
"What is the time right now?"
"I am very happy today."
"What is the weather like?"
"The quick brown fox jumps over the lazy dog."
STOP RECOGNITION
Step 2: Recognize with TV background noise

Extending VoiceFilter-Lite to multiple users

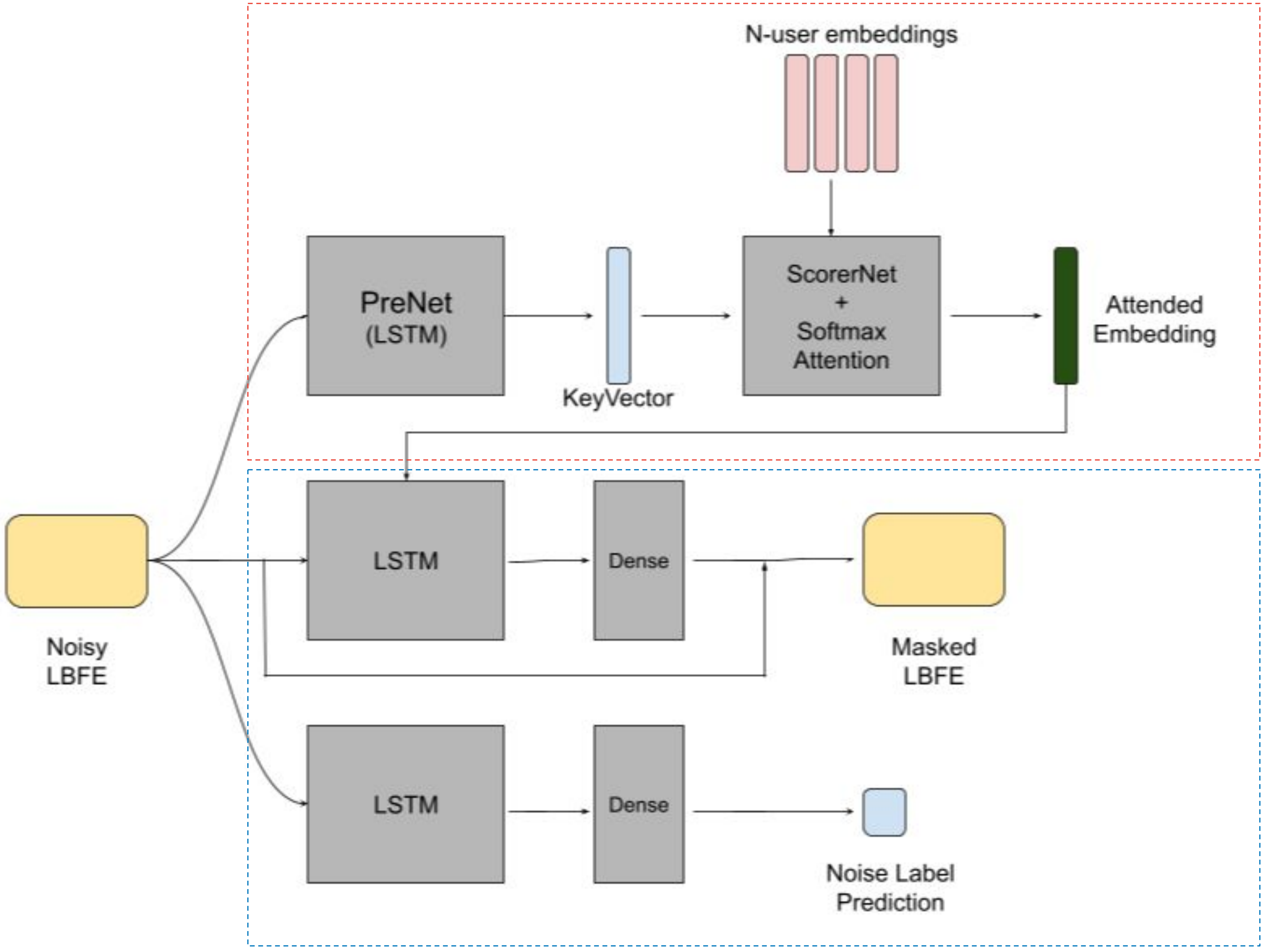
- Smart home speakers are **shared devices**
- Most households have multiple family members
- It is important to extend VoiceFilter-Lite to **multiple enrolled users**

Options for a multi-user VoiceFilter-Lite:

1. **Multiple passes of the same VoiceFilter-Lite model - once for each speaker**
   - × Computationally inefficient to run multiple passes of the model on-device
   - × Infeasible: requires complex logic to select the best output from each pass
   - × Memory intensive
2. **A single VoiceFilter-Lite model that uses all embedding inputs**
   - × The order of the concatenated embedding inputs matters (not permutation invariant)
3. **A single VoiceFilter-Lite model that uses attention to select the target speaker**
   - ✓ Computationally more efficient
   - ✓ Permutation invariant
   - ✓ Supports an arbitrary number of enrolled users in a single pass

Multi-user VoiceFilter-Lite (MUVF) model Architecture

Model size: 3.3 MB



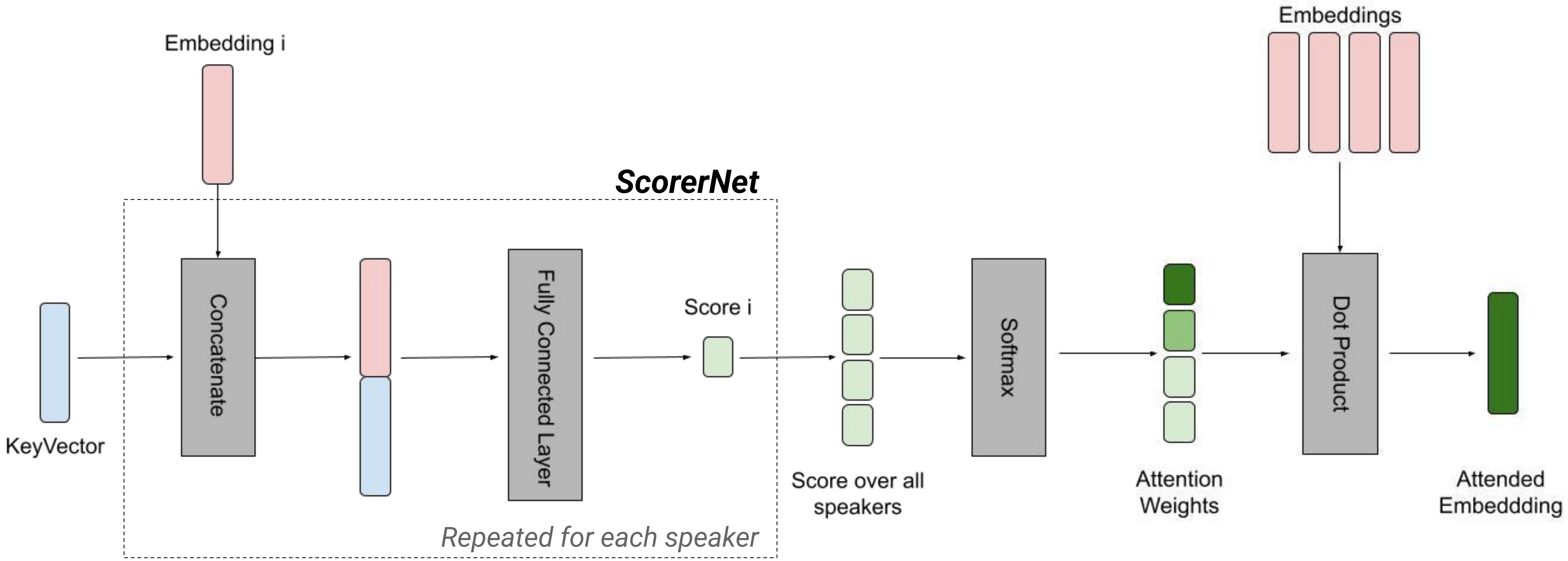
AttentionNet

VoiceFilterNet

(identical to single-user VoiceFilter-Lite)

MUVF uses attention to compute the ***most likely* target speaker embedding** from the input conditioned on a set of known speaker profiles

AttentionNet Architecture



- The **ScorerNet** computes a similarity score between the KeyVector and each of the speaker embeddings and outputs a set of N attention weights
- The **Attended Embedding** is the dot product of the weights and the embedding inputs

AttentionNet and VoiceFilterNet are trained in an end-to-end manner

$$\boxed{\mathrm{L}_{\mathrm{total}} = w_1 \mathrm{L}_{\mathrm{asym}} + w_2 \mathrm{L}_{\mathrm{noise}} + w_3 L_{att}}$$

Asymmetric reconstruction loss - ensures that the enhanced Spectrogram matches the clean spectrogram (Ground Truth)

$$\mathrm{L}_{\mathrm{asym}} = \sum_t \sum_f \left(g_{asym}\left(S_{clean}(t,f) - S_{enh}(t,f), \alpha\right)\right)^2$$

Noise label prediction loss - ensures that predicted noise label is close to the ground truth label

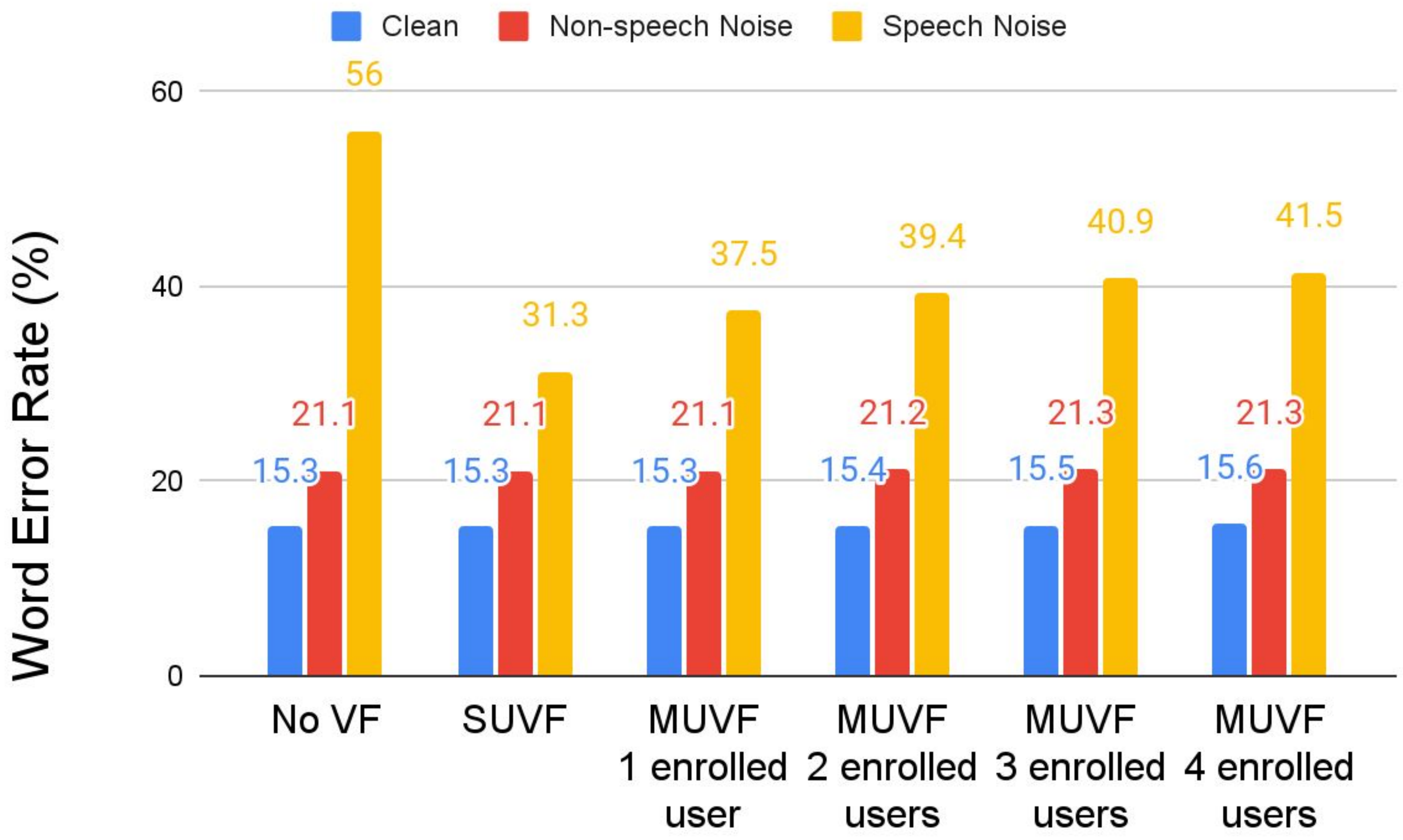
$$\mathrm{L}_{\mathrm{noise}} = \sum_i (n_{pred} - n_{gt})^2$$

Attention loss - minimizes the mean squared error between the attended embedding and the ground truth embedding from the target speaker.

$$\mathrm{L}_{\mathrm{att}} = \sum_t \left\| e_{att}^{(t)} - e_{gt} \right\|^2$$

MUVF → ASR improves Word Error Rate compared to no VF

Experiment 1: Speech recognition task under various noise conditions.

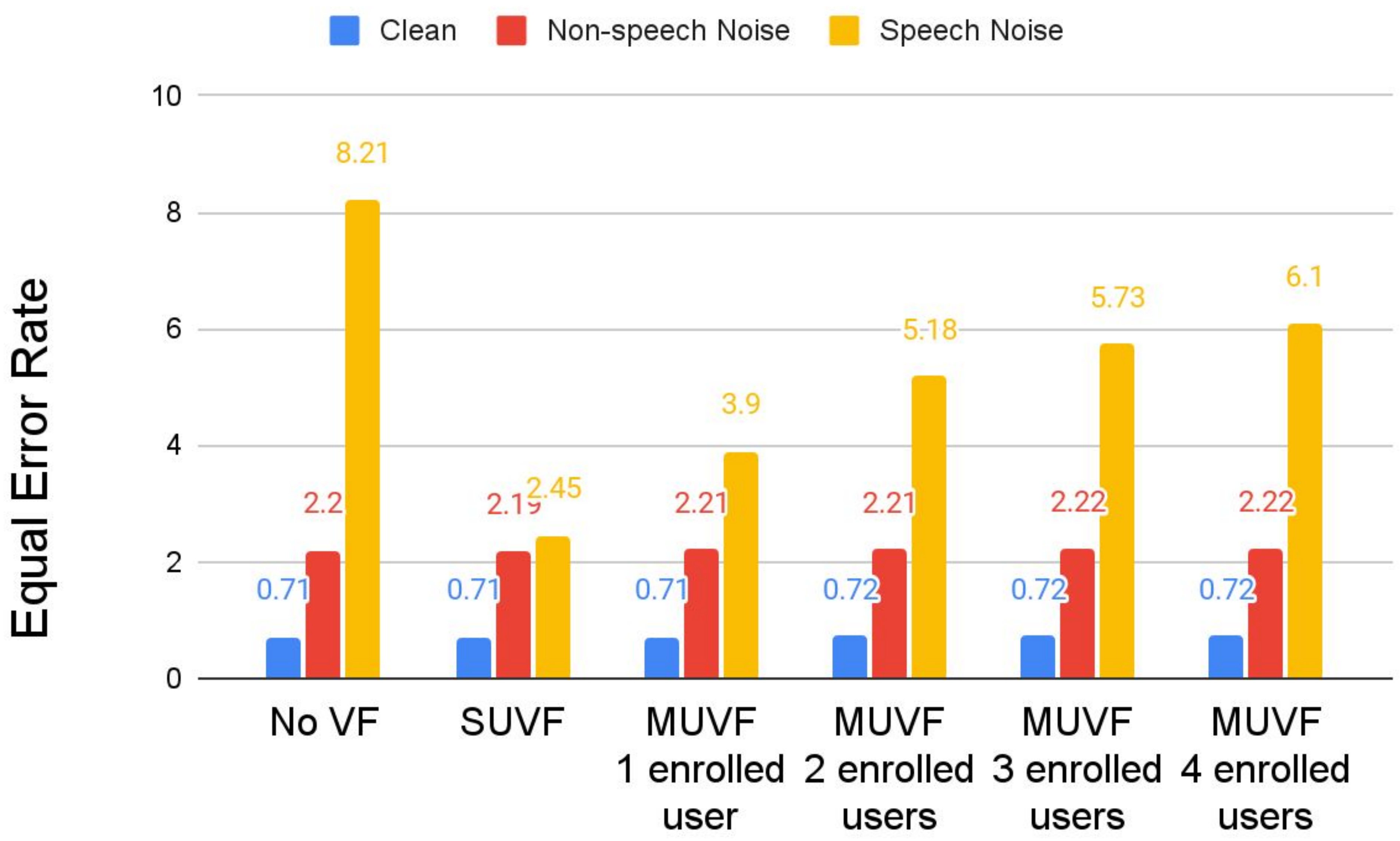


Vendor-collected dataset (230 speakers, 20K utterances)

- MUVF was placed in the feature frontend of an on-device, streaming ASR model
- Relative to no VF, **MUVF with 4 enrolled users** decreases WER by **25.9%**
- Enrolling more speakers degrades performance since selecting the correct speaker from overlapping speech is a difficult task

MUVF → TI-SV improves speaker verification accuracy compared to no VF

Experiment 2: Speaker Verification task under various noise conditions.



Vendor-collected dataset (958 speakers, 220K utterances)

Note: Only SNR 0dB, additive noise condition is shown

- MUVF was placed in the feature frontend of an on-device Text-independent Speaker Verification model
- Relative to No VF, **MUVF with 4 enrolled users (MUVF-4)** reduces the EER by **25.7%**
- Enrolling more speakers degrades performance since selecting the correct speaker from overlapping speech is a difficult task

Application of multi-user VoiceFilter-Lite:

Personalized keyphrase detection

Allow users to say ***specific keyphrases*** to smart devices without the wake word

Comment: OK Google, I'm exhausted saying 'Google'

Stephen Hall - May. 18th 2020 1:21 pm PT @hallstephenj

https://9to5google.com/2020/05/18/comment-ok-google-im-exhausted-saying-google/

People who want to have (more) real conversations with their speaker bot.

Where Google really shows its intelligence is its ability to understand contextual questions.

https://www.buzzfeednews.com/article/nicolenguyen/google-home-review

Avoiding the *wake word* would make interactions with the smart device more natural

Detecting keyphrases in the ambient environment is challenging

Challenge 1: False Triggering by ambient speech

Ambient speech, from a TV or family members in the room can false trigger the device.

Proposed Solution: Responding to known / enrolled speakers via Speaker Verification

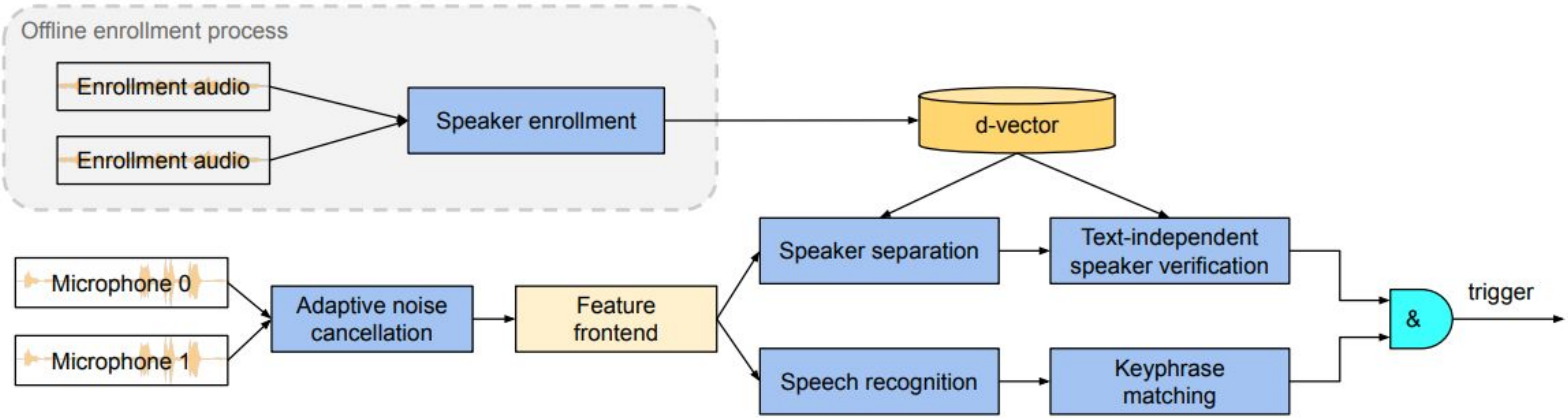
Detecting keyphrases in the ambient environment is challenging

Challenge 2: False Rejection by ambient speech

Overlapping speech can make speaker identification less accurate.

Proposed Solution: Identify and suppress overlapping speech via *VoiceFilter-Lite*

Proposed personalized keyphrase detector system



A query is valid if the following two conditions are met [2]:

1. The ASR model recognizes the keyphrase
2. The Speaker Verification model recognizes the speaker as an enrolled user

[2] R. Rikhye, Q. Wang, et al., "Personalized Keyphrase Detection using Speaker and Environment Information" in Proc. Interspeech, 2021

Speaker Verification increases False Rejects when there is ambient speech

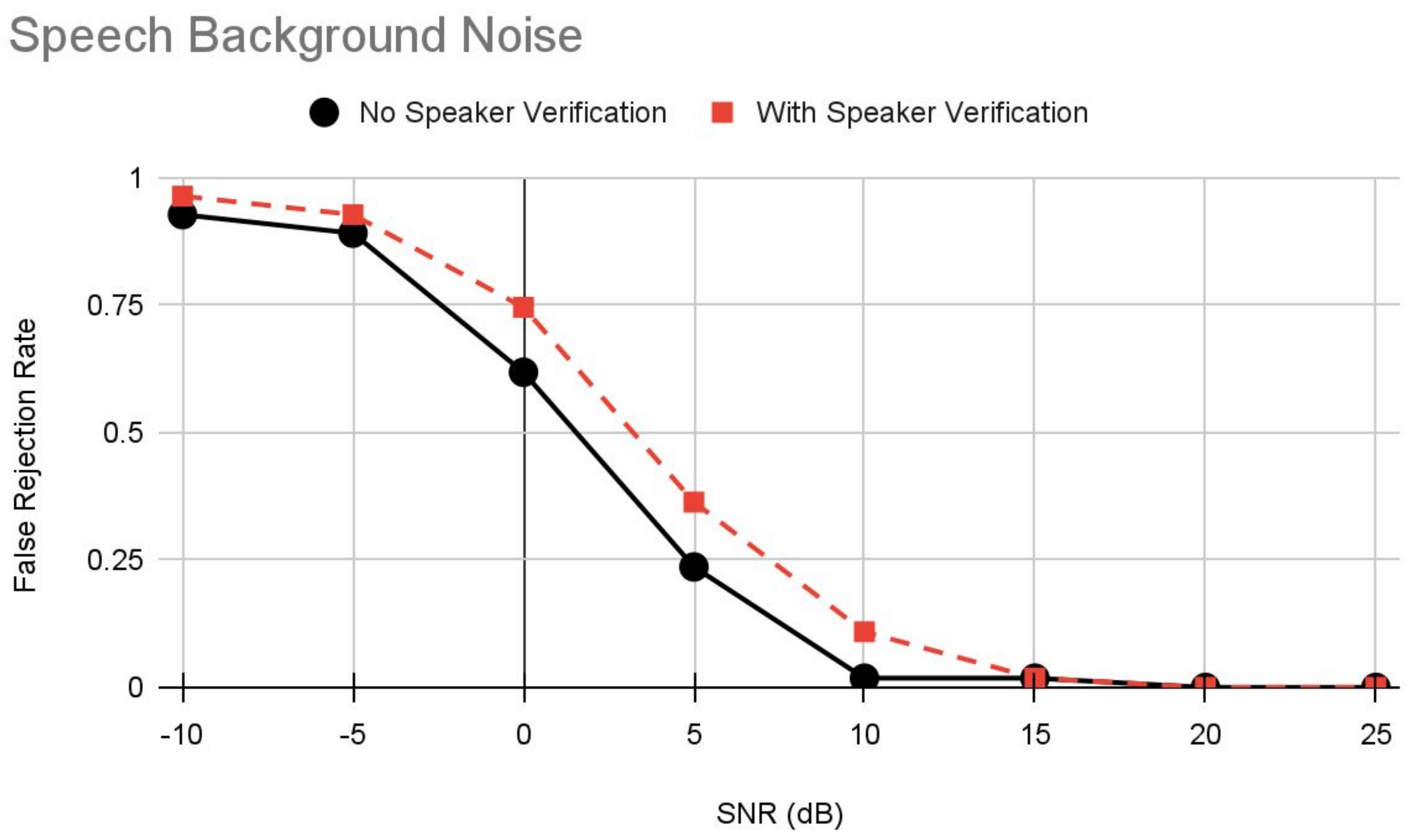
YouTube dataset with no queries (300 hours)

False accepts* per hour

Without TI-SV	With TI-SV (4 enrolled speakers)
0.2746	0.03457 (**-91.7%**)

*False accept = query that is wrongly accepted as a keyphrase

Vendor-collected dataset (303 speakers, 92K queries, 97 hours)



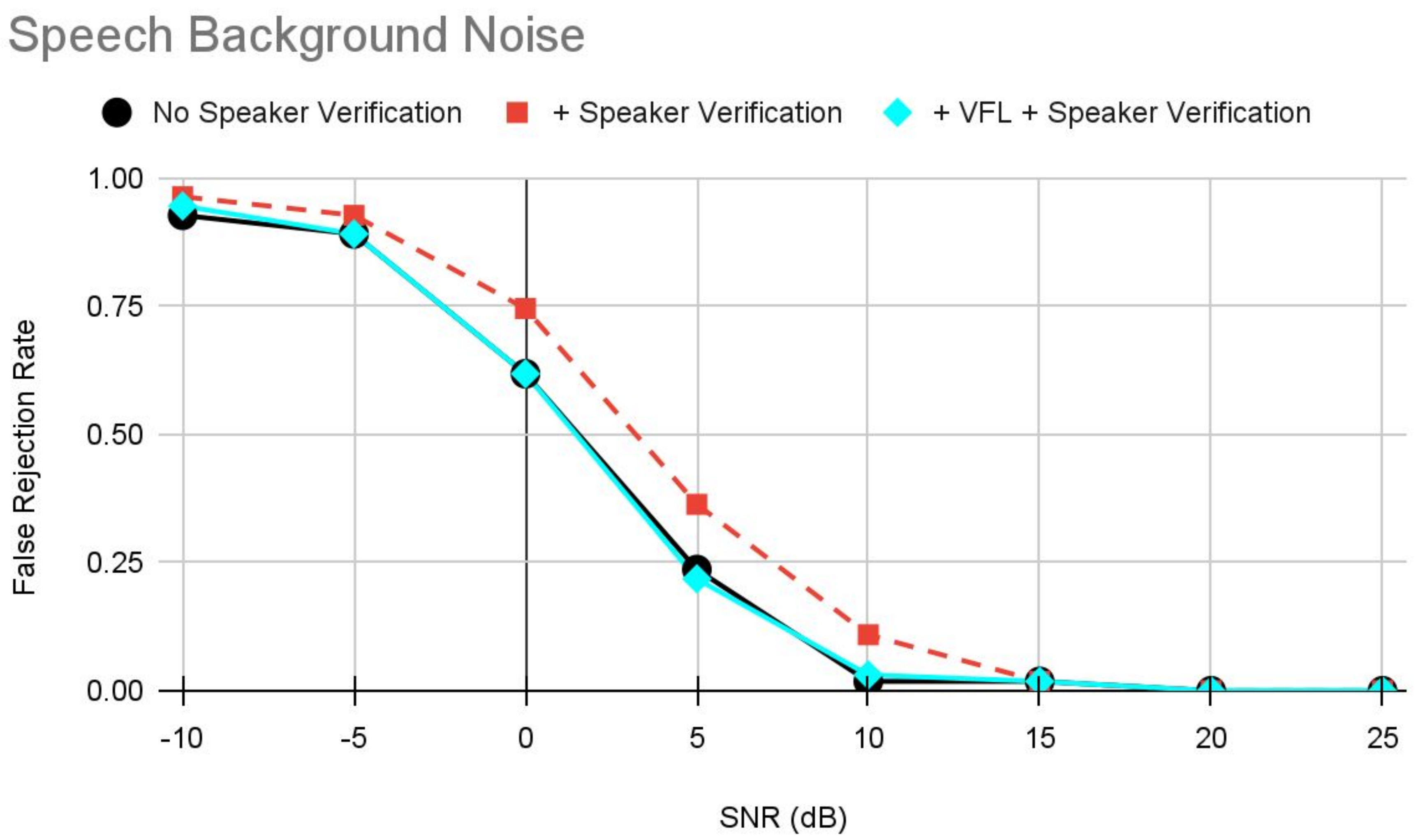
*False reject = valid keyphrase that is wrongly rejected

- Adding TI-SV ***significantly reduces*** the number of False Accepts per hour
- Adding TI-SV ***increases*** the False Reject Rate when there is overlapping speech
- A major source of speaker verification False Rejects is **multi-talker speech**

VFLite → TI-SV increases speaker identification accuracy and reduces False Rejects

Vendor-collected dataset (303 speakers, 92K queries, 97 hours)

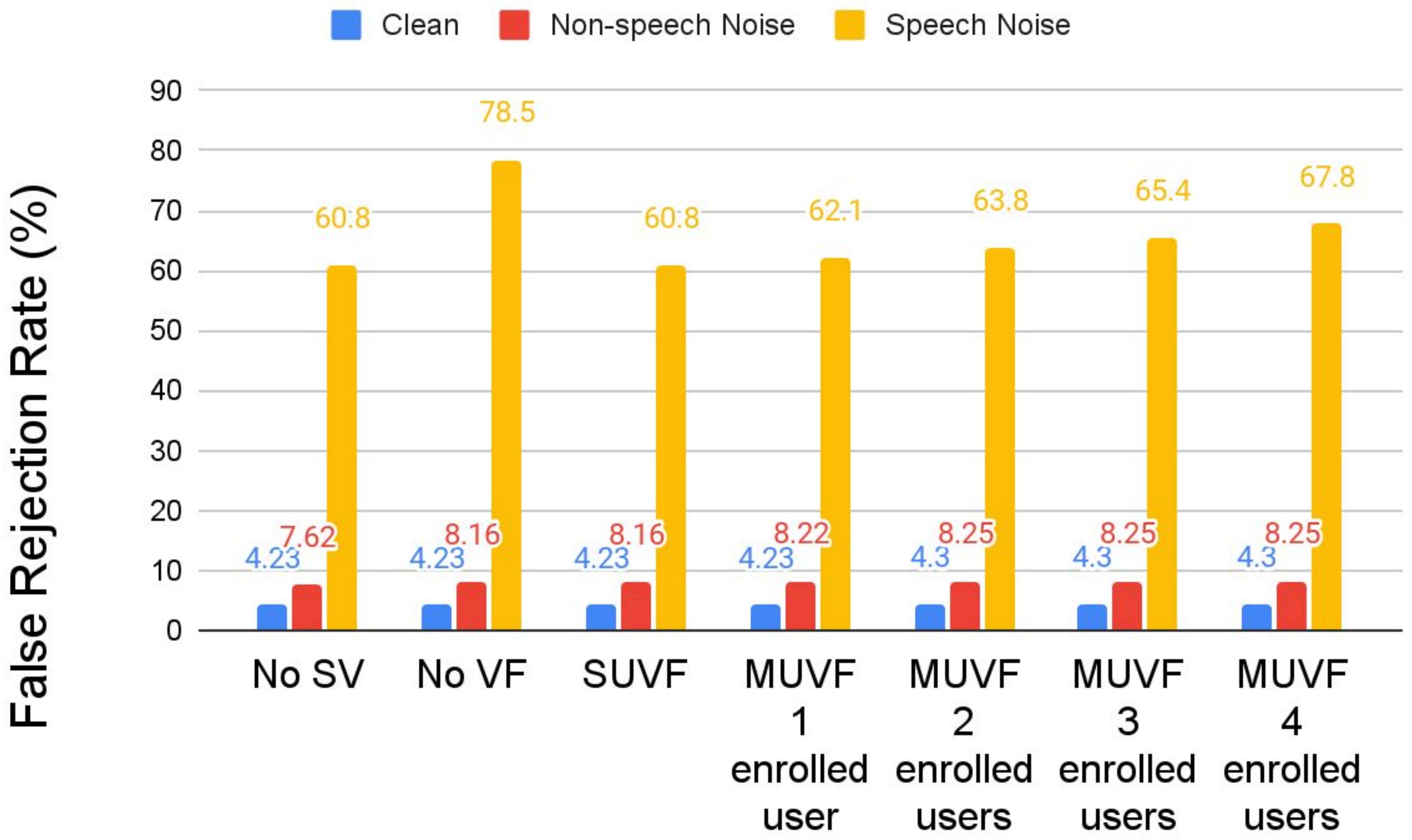
Noise source	Room	SNR (dB)	EER (%)	
			No VFL	With VFL
Speech	Additive	-5	12.83	**4.24**
		0	8.34	**2.35**
		5	4.99	**1.47**
	Reverb	-5	17.76	**7.03**
		0	11.04	**3.63**
		5	6.41	**2.09**



- VF-Lite → TI-SV results in a **~67%** improvement in speaker identification EER
- This mitigates speaker identification errors during overlapping speech
- However this is only limited devices with a single enrolled user

With VF-Lite, we prevent the increase in False Rejects with ambient speech!

MUVF → TI-SV also reduces False Rejects



Vendor-collected dataset (303 speakers, 92K utterances)

Note: Only SNR 0dB, additive noise condition is shown

- Relative to No VF, MUVF with 4 enrolled users reduces the FRR by **13.8%**
- The reduction in FRR is worse than the SUVF since selecting the correct speaker under ambient noise conditions is fundamentally a more challenging task.

Summary

- A novel **attention mechanism** identifies which of the *N* enrolled users is speaking in a particular frame.

Summary

- A novel **attention mechanism** identifies which of the *N* enrolled users is speaking in a particular frame.
- This **attentive embedding** can then be used with any speaker condition speech model like VoiceFilter-Lite, Personal Voice Activity Detection, or Personalized ASR.

Summary

- A novel **attention mechanism** identifies which of the *N* enrolled users is speaking in a particular frame.
- This **attentive embedding** can then be used with any speaker condition speech model like VoiceFilter-Lite, Personal Voice Activity Detection, or Personalized ASR.
- In the multi-user VoiceFilter-Lite application, we show that with up to 4 enrolled users, relative to no VF and in the presence of overlapping speech background noise, MUVF is able to:
  - **Improve** speaker verification accuracy
  - **Reduce** Word Error Rate
  - **Reduce** keyphrase False Rejection Rate

Summary

- A novel **attention mechanism** identifies which of the *N* enrolled users is speaking in a particular frame.
- This **attentive embedding** can then be used with any speaker condition speech model like VoiceFilter-Lite, Personal Voice Activity Detection, or Personalized ASR.
- In the multi-user VoiceFilter-Lite application, we show that with up to 4 enrolled users, relative to no VF and in the presence of overlapping speech background noise, MUVF is able to:
    - **Improve** speaker verification accuracy
    - **Reduce** Word Error Rate
    - **Reduce** keyphrase False Rejection Rate
- We observe a degradation in performance with more enrolled users. This is because the AttentionNet has a difficult task of selecting the correct speaker from noisy input.
    - Our future work aims at addressing this discrepancy

Thank you.

Questions?

Google