Google

Personalized Keyphrase Detection using Speaker and Environment Information

Authors: ***Rajeev Rikhye, Quan Wang***, *Qiao Liang, Yanzhang He, Ding Zhao, Yiteng (Arden) Huang, Arun Narayanan, Ian McGraw*

30 August – 3 September 2021
INTERSPEECH 2021 BRNO | CZECHIA
Speech everywhere!

Abstract

We introduce a **keyphrase detection** system:

- **Customizable**: Can detect any phrase composed of words from a large vocabulary
- **Streaming** inference: Using RNN-T ASR model
- **On-device**: Pruning the model
- **Personalized**: Using a text-independent speaker verification model
- **Noise-robust**:
  - **Multi-microphone**: Adaptive noise cancellation (ANC) with *Speech Cleaner*
  - **Multi-talker**: *VoiceFilter-Lite* for speaker verification

System diagram

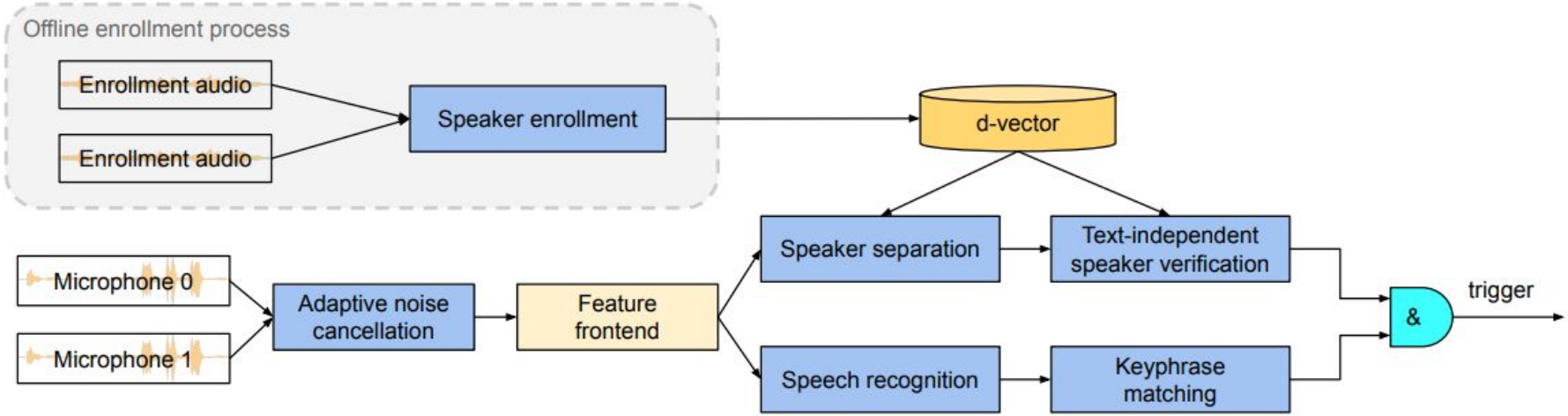


Figure 1: *Diagram of the proposed keyphrase detection system. The d-vector is obtained in a separate offline enrollment process.*

Key results

Text-independent **speaker verification**:

- Reduces FA/hour by rel. 91%
- But it increases FRR by rel. 20.6% in the multi-talker scenario

Speaker separation (*VoiceFilter-Lite*) in multi-talker scenario:

- Reduces speaker verification EER by rel. 67.4%
- Reduces keyphrase detection FRR by rel. 29.4%

Adaptive noise cancellation (*Speech Cleaner*):

- Reduces FRR by rel. 68.3% in the non-speech noise
- Reduces FRR by rel. 25.2% in the multi-talker scenario

Before - Slides for 3min video

After - Slides for 15min video

Google

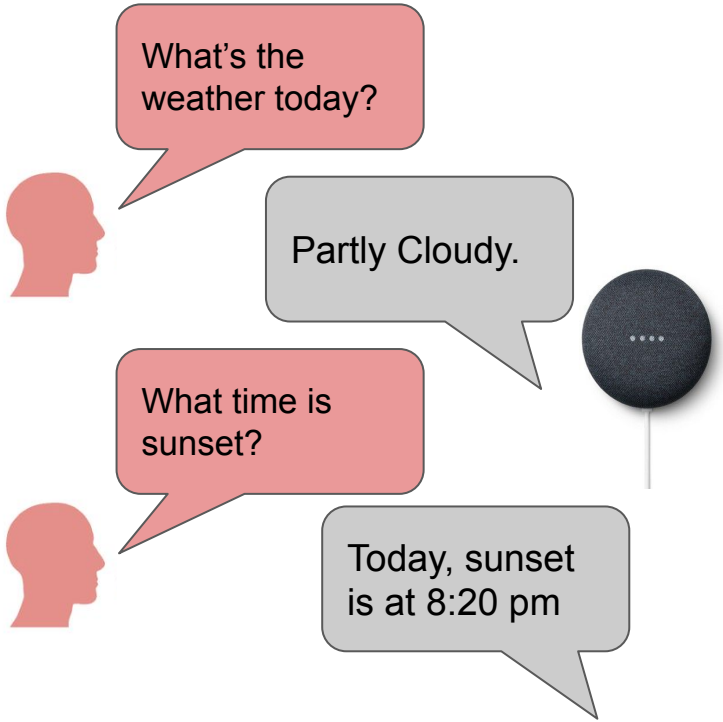
Personalized Keyphrase Detection using Speaker and Environment Information

Authors: ***Rajeev Rikhye, Quan Wang***, *Qiao Liang, Yanzhang He, Ding Zhao, Yiteng (Arden) Huang, Arun Narayanan, Ian McGraw*

30 August – 3 September 2021
INTERSPEECH 2021 BRNO | CZECHIA
Speech everywhere!

Interactions with smart devices currently require a *wake word*

Interactions with smart devices currently require a *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 more naturalistic.

Our goal is to allow users to say ***specific keyphrases*** to smart devices without requiring the wake word

We introduce a **keyphrase detection** system:

- **Customizable**: Can detect any phrase composed of words from a large vocabulary
- **Streaming** inference: Using RNN-T ASR model
- **On-device**: Pruning the model to meet on-device memory constraints
- **Personalized**: Using a text-independent speaker verification model
- **Noise-robust**:
  - **Multi-microphone**: Adaptive noise cancellation (ANC) with *Speech Cleaner*
  - **Multi-talker**: *VoiceFilter-Lite* for speaker verification

Detecting keyphrases is challenging

Example keyphrases

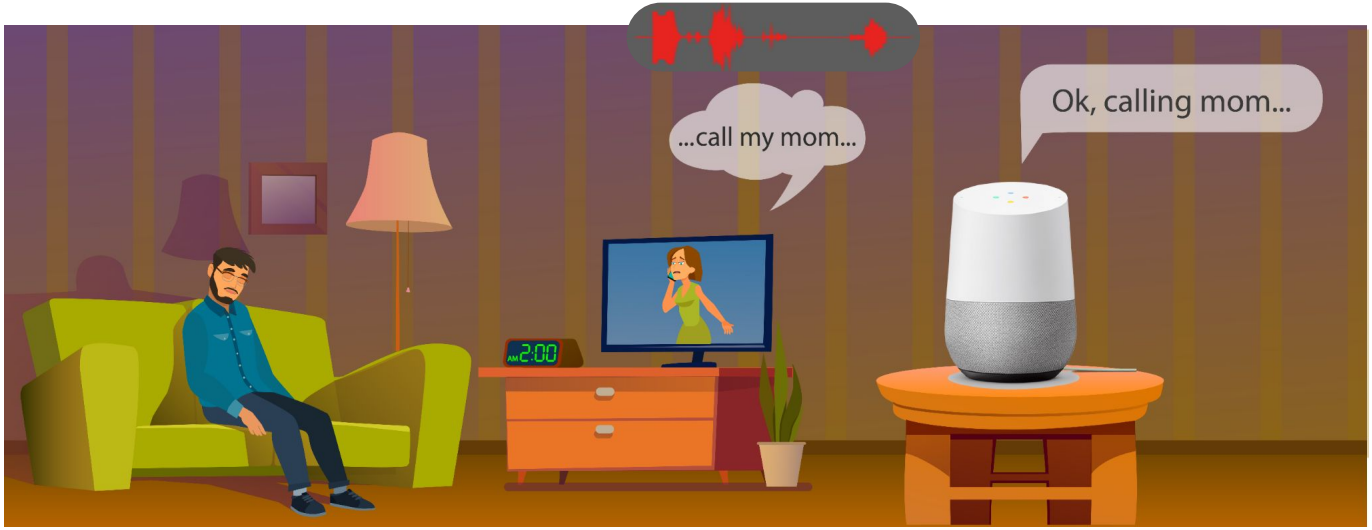
“Turn on the lights”, *“Stop the music”*, *“Set an alarm for 5 a.m.”*

1. Must be able to detect a large corpus of keyphrases.
2. Keyphrases may have variable length and audio durations
   - A single word: *“Stop”*
   - Sentences: *“Play my workout playlist”*
3. The set of recognized keyphrases should be customizable without requiring training or new models.

Our system is a generic ASR model that allows user-defined keyphrases, providing greater flexibility to the end users

Detecting keyphrases in a ***noisy environment*** is challenging

Challenge 1: False Triggering by ambient speech



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

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

Challenge 2: False Triggering by ambient noise

General ambient noise (eg. music, barking dog) can also false trigger a query.

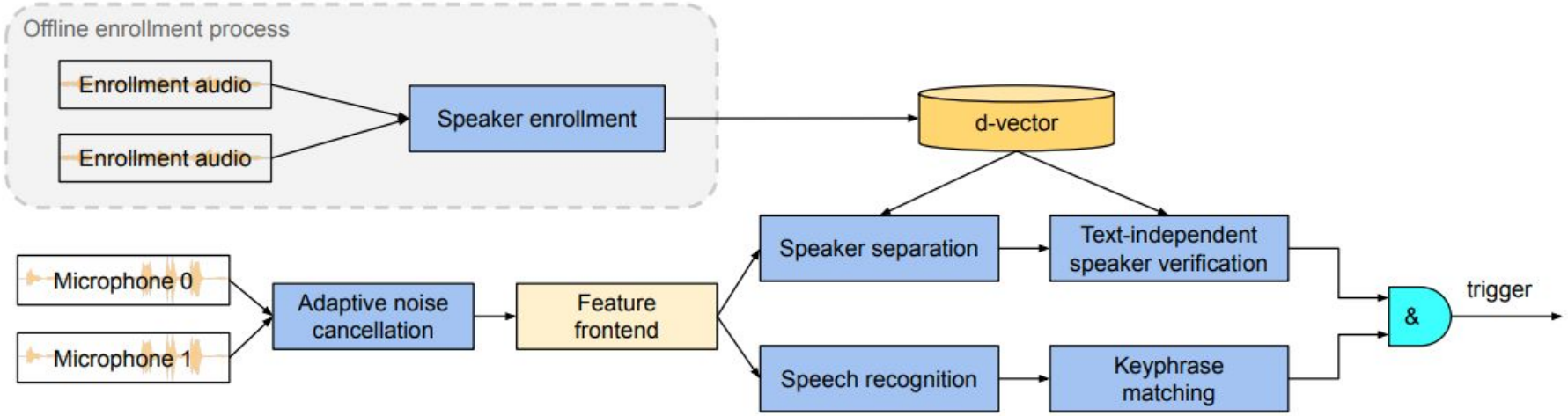
Proposed Solution: Suppress background noise via *Speech Cleaner*

Challenge 3: False Rejection by ambient speech

Overlapping speech can make speaker identification less accurate.

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

Proposed Keyphrase detection system



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

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

Speaker verification ***significantly reduces*** false triggering

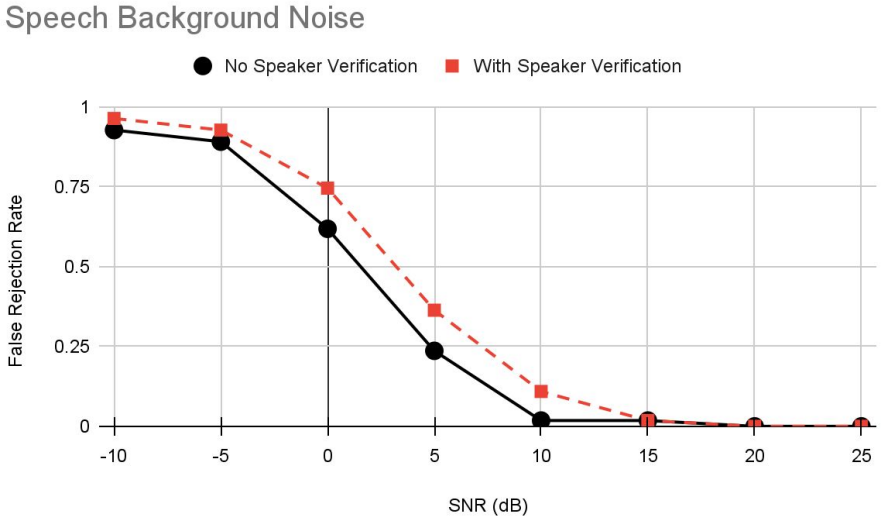
False accepts* per hour

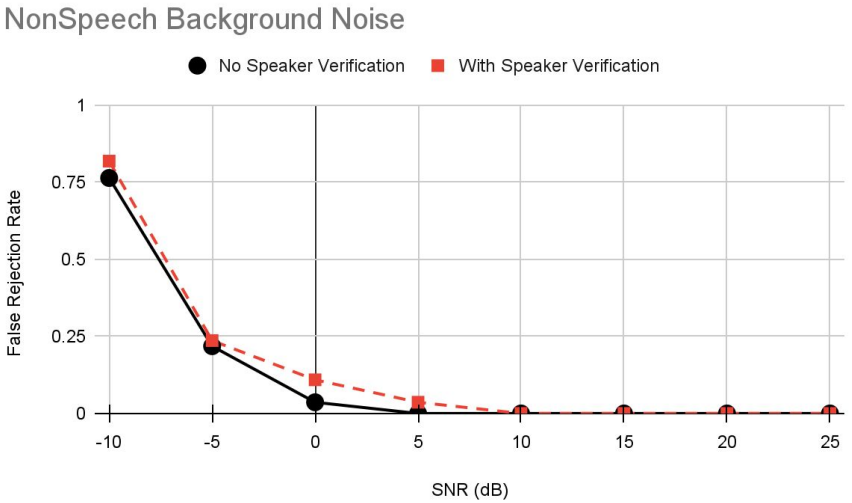
	Without TI-SV	With TI-SV. (5 enrolled speakers)	With TI-SV (1 enrolled speaker)
YouTube dataset (with no queries)	0.2746	0.03457 (**-91.7%**)	0.00985 (**-97.5%**)

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

- We used Text-independent Speaker Verification (TI-SV)
- Queries that do not match the enrolled user(s) are rejected.
- Most TV/radio background speech will be rejected by speaker verification alone.

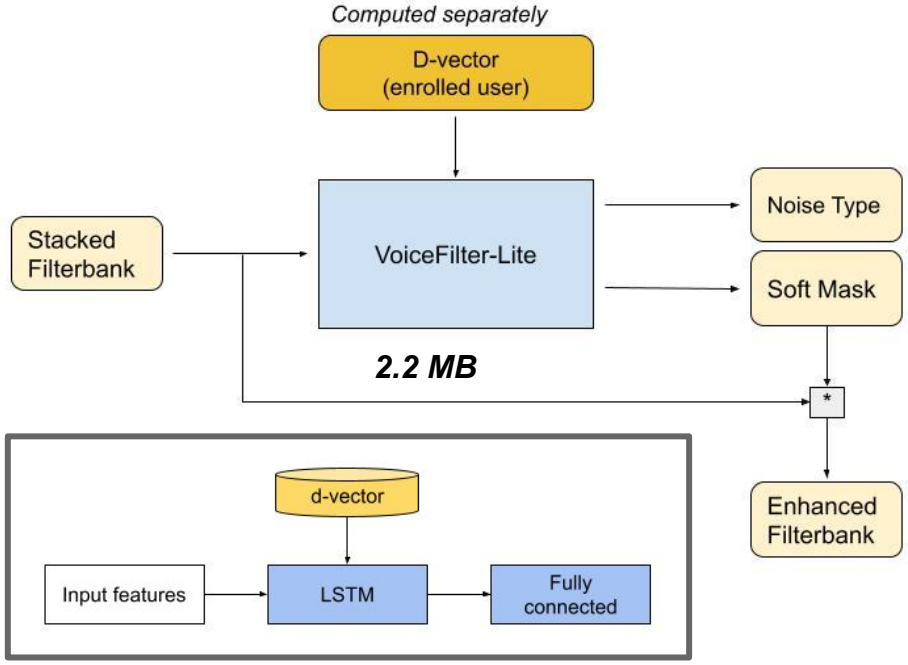
Speaker Verification ***increases false rejections*** in the presence of overlapping speech





- A major source of error is enrolled speaker mis-identification
- Overlapping speech masks features making it harder to recognize the speaker
- This is a common problem for speaker verification

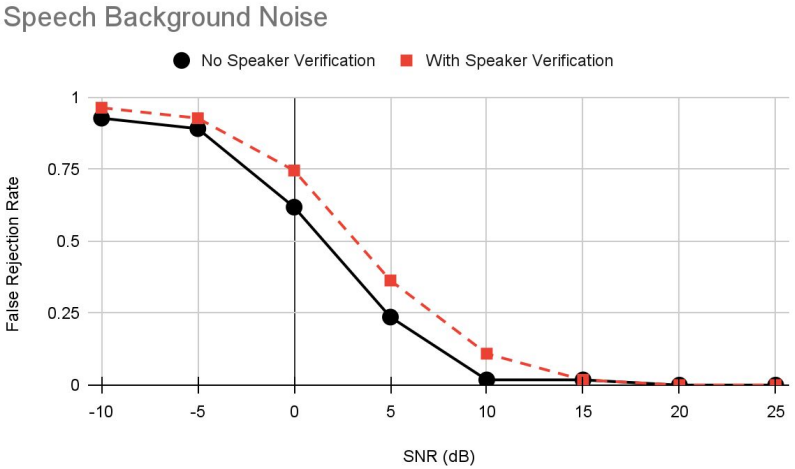
VoiceFilter-Lite enhances enrolled speaker features from overlapping speech



- In overlapping speech:
  - Enrolled user features are enhanced.
  - Non-enrolled features are suppressed.
- Non-overlapping speech:
  - Filterbank is not modified
- VoiceFilter-Lite was applied to the ASR frontend.

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

VoiceFilter-Lite can improve speaker recognition in multi-talker scenarios



Improvements

- VF-Lite can suppress frames with overlapping speech before SV
  - This will improve the speaker verification accuracy
- VF-Lite now supports multiple users.

Hypothesis:

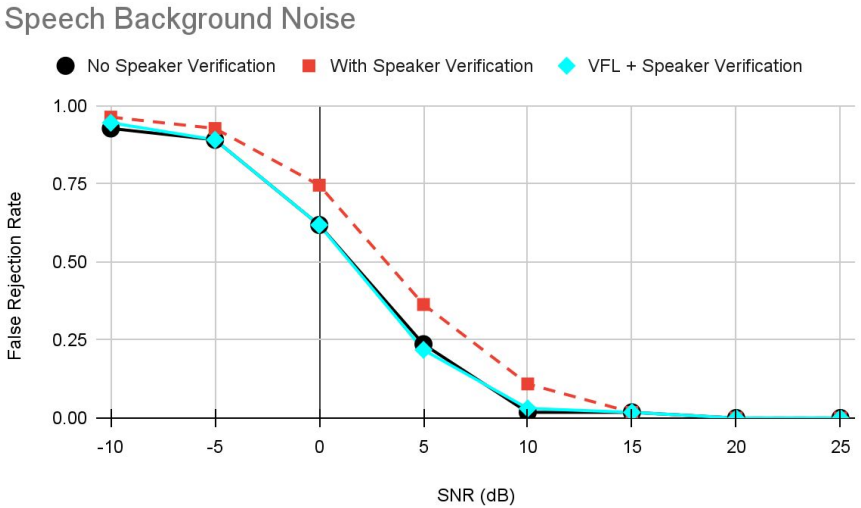
Adding VoiceFilter-Lite to the SV frontend (instead of ASR) will help to suppress overlapping speech and improve speaker verification accuracy

Google

R. Rikhye, et al., "Multi-user VoiceFilter-Lite via attentive speaker emebdding" ArXiv:2107.01201

VF-Lite → SV increase speaker identification accuracy and reduces False Rejects during overlapping speech

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**

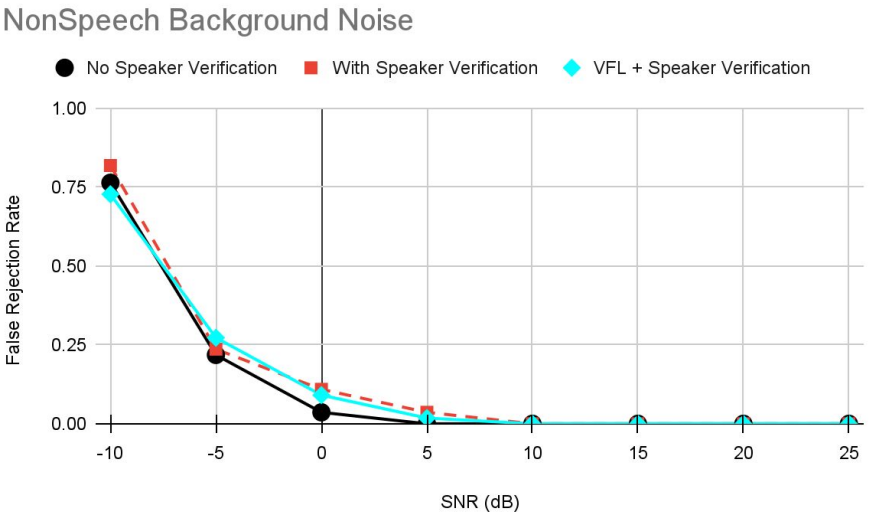


- VF-Lite → TI-SV results in a **~67%** improvement in speaker identification EER.
- This mitigates speaker identification errors during overlapping speech.

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

VF-Lite → SV does not have an effect when there is no overlapping speech.

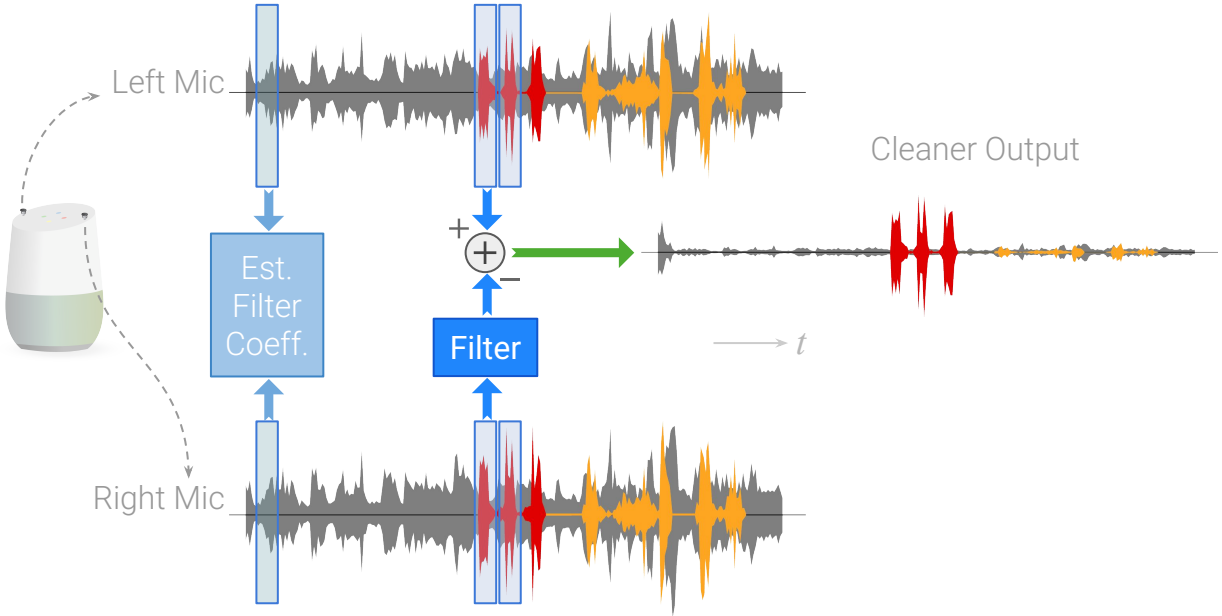
Noise source	Room	SNR (dB)	EER (%)	
			No VFL	With VFL
Clean			0.65	0.64
Non-speech	Additive	-5	5.30	5.23
		0	2.04	2.01
		5	1.22	1.22
	Reverb	-5	6.51	6.53
		0	2.90	2.91
		5	1.60	1.59



- VF-Lite passes the filterbank through unmodified in non-overlapping speech
- No change in EER or FR when background noise does not contain speech

We require another solution to make the system robust to non-speech background noise

Hotword Cleaner was developed to enhance wake word detection in noisy environments

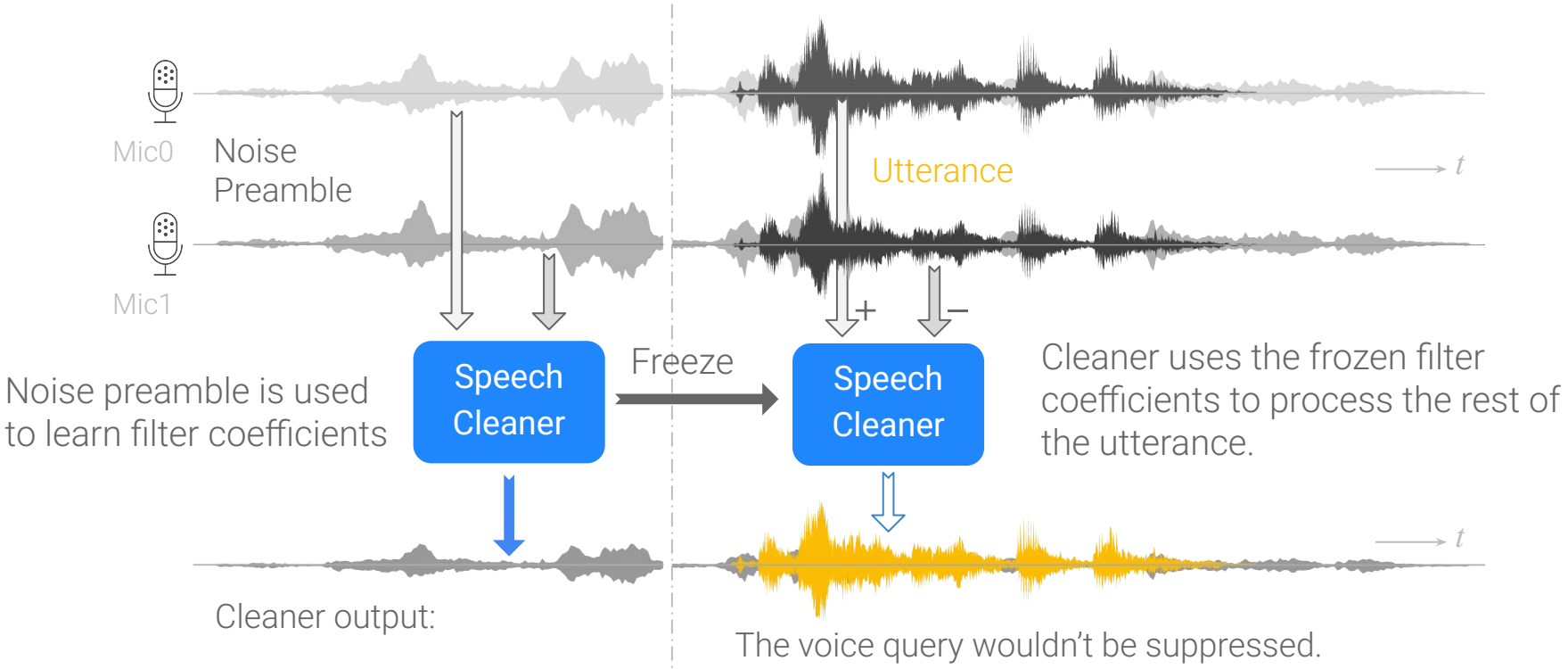


- Most smart devices have 2 or more microphones
- Filter coeffs are computed from the noise preamble and applied to the hotword

Google

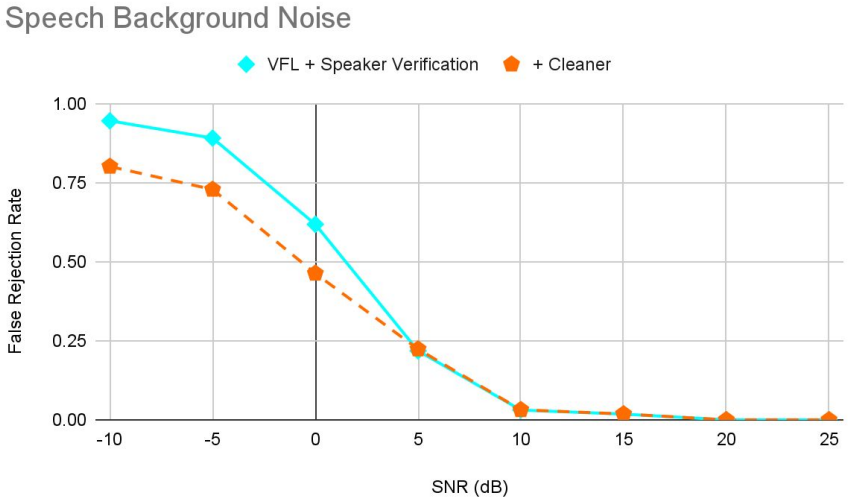
A. Huang, et al., "Hotword cleaner: dual-microphone adaptive noise cancellation with deferred filter coefficients for robust keyword spotting," in ICASSP. IEEE, 2019, pp. 6346–6350.

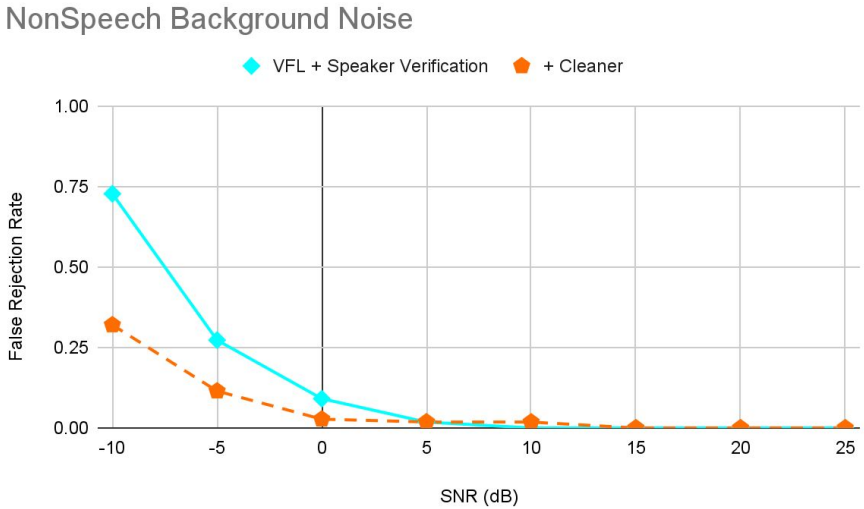
Speech Cleaner learns and suppresses background noise during the query



We add *Speech Cleaner* to the frontend of both Speaker Verification and ASR

Speech Cleaner helps to further ***reduce*** False Rejection Rate





- Speech Cleaner reduces ASR errors, allowing the system to detect keyphrases in the presence of either speech or non-speech noise

Summary

- We introduce a **personalized keyphrase detection** system that is highly customizable and robust to different types of background noise
- This keyphrase detection system allows users to interact with their smart devices without having to say a wake word
- We leveraged speaker and environment information to reduce both false triggering and false rejections to improve user experience

Summary

Challenge 1: False Triggering by ambient speech

- Speaker Verification (***TI-SV***) helps to reduce false accepts by ~91% (relative to model with no SV) by rejecting queries that are not from the enrolled user(s))

Summary

Challenge 1: False Triggering by ambient speech

- Speaker Verification (***TI-SV***) helps to reduce false accepts by ~91% (relative to model with no SV) by rejecting queries that are not from the enrolled user(s)

Challenge 2: False Triggering by background noise

- Adaptive noise cancellation (***Speech Cleaner***) helps to suppress both speech and non-speech background noise, improving keyphrase detection by ~68%

Summary

Challenge 1: False Triggering by ambient speech

- Speaker Verification (***TI-SV***) helps to reduce false accepts by ~91% (relative to model with no SV) by rejecting queries that are not from the enrolled user(s)

Challenge 2: False Triggering by background noise

- Adaptive noise cancellation (***Speech Cleaner***) helps to suppress both speech and non-speech background noise, improving keyphrase detection by ~68%

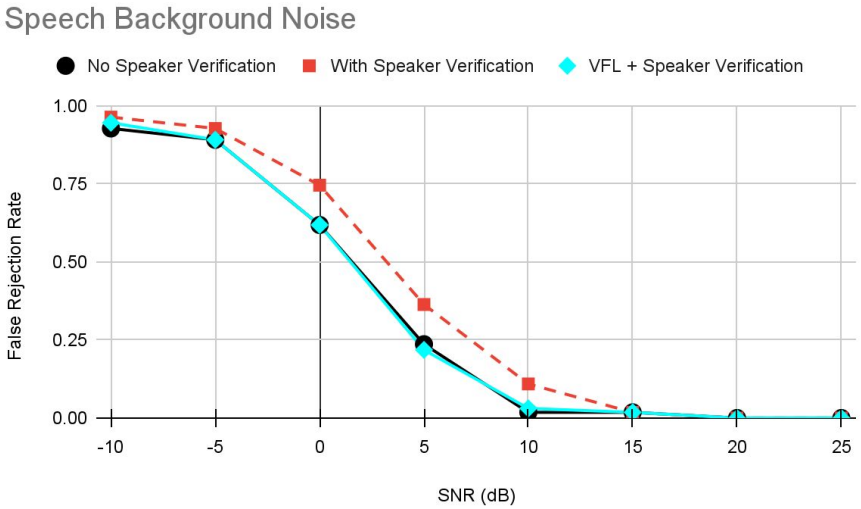
Challenge 3: False rejection by ambient speech

- Speaker separation (***VoiceFilter-Lite***) in the TI-SV feature frontend suppresses overlapping speech and improves FRR by ~29% (relative to model with TI-SV)

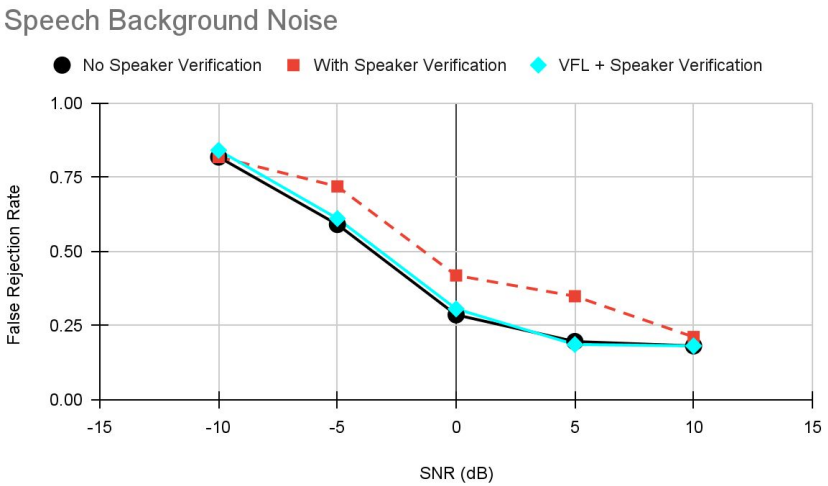
Supplementary Slides

VF-Lite → TI-SID reduces False Rejects during overlapping speech

Vendor (Webhound) data



Multi-speaker TTS dataset



With VF-Lite we can reduce FA without an increase in FR with ambient speech!