OPTIMIZE FOR MY VOICE WITH SPEAKER IDENTIFICATION

Marcin Ciołek, Michał Sulewski, Rafał Pilarczyk, Raul Casas, Samer Hijazi, Scott Plude, Michelle Mao, Guoqing Zhang, Nathan Rickey, Mahesh Godavarti, Kamil Wójcicki, Ali Mouline, Savita Kini, Marta Chełkowska, Taha Emara, Yusuf Isik, Amir Abdelwahed

Cisco Systems, Inc., Babblelabs team
marciole@cisco.com, msulewsk@cisco.com

Abstract

The proposed system enhances speech in Cisco Webex video-conferencing applications. The demo aims to preserve the primary talker while suppressing interfering talkers, noise, and reverberation. Besides these challenges, the system automatically controls the volume of the primary talker. The novelty of the proposed system is given by implementing adaptive primary talker detection and tracking while preserving fast and accurate far-field talker attenuation.

1. Webex conferencing platform

I. Speech enhancement modes

Webex Smart Audio allows users to choose one of the following options:

A) **Noise Removal** – removes all background noise

B) **Optimize for my Voice** – removes all background noise and background speech

C) **Optimize for all Voices** – removes all background noise and enhances all voices

D) **Music Mode** – others hear the original sound when you play an instrument or sing nearby

II. Optimize for my Voice

In order for this mode to work properly some requirements need to be fullfilled:

A) **primary talker** – stays up to 1 m from the mic

B) **interfering talker** – reverberated background speech

C) **target volume** for a primary talker: -26 dBrms

III. Problem statement

1) In speakerphone mode, primary talker might be chopped/suppressed if he moves farther than 1 m

2) Word chopping is caused by temporarily misdetecting a primary talker as an interfering talker

IV. Demo goals

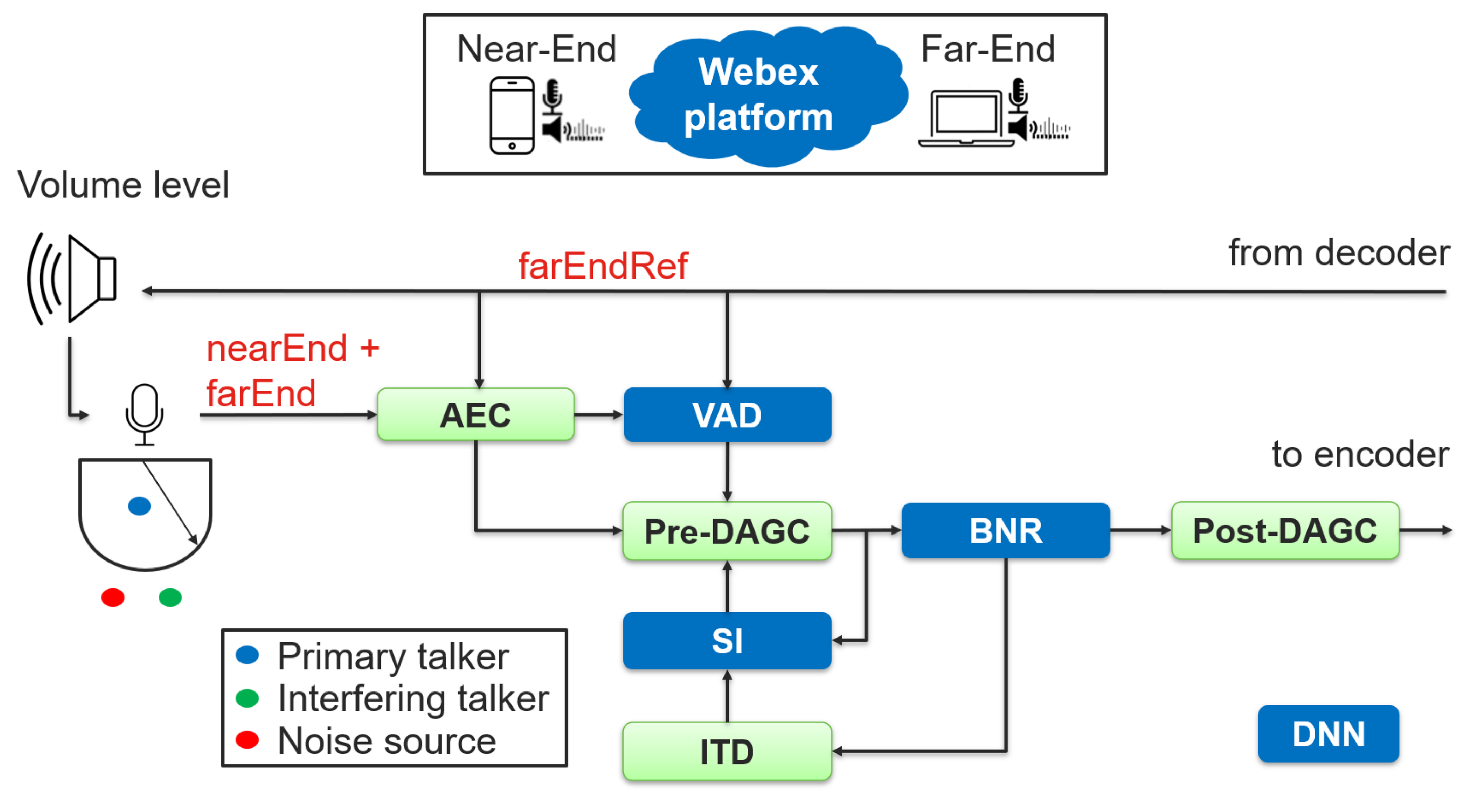
Speaker identification can help with:

1) preserving and levelizing the primary talker's speech at farther distances

2) reducing word chopping

Distance [m]	No Speaker ID	Speaker ID
Silent room	Primary talker	
< 1	preserved	preserved
1 - 2	word chopping	preserved
2 - 3	suppressed	word chopping
> 3	suppressed	suppressed

2. System diagram



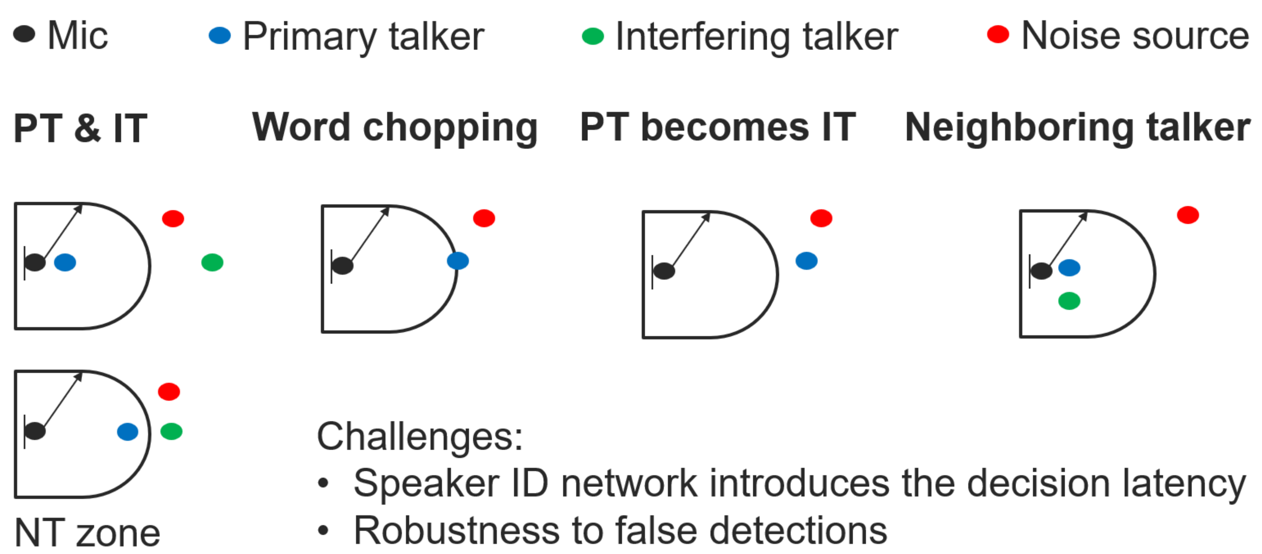
AEC – Acoustic Echo Cancellation
VAD – Voice Activity Detection
BNR – Background Noise Removal
DAGC – Digital Adaptive Gain Control
ITD – Interfering Talker Detection
SI – Speaker Identification

3. Speaker identification

I. Network specification

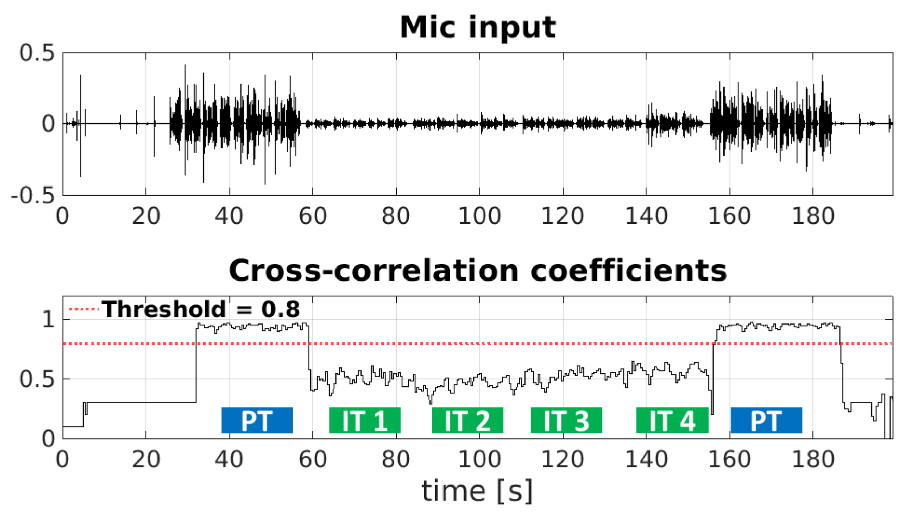
- Network architecture: Mobilenetv2-like Architecture
- Network trained on 850 speakers from Common Voice
- Recordings: 1.7 million
- Cost function: Arcface
- Network Receptive Field: 2s (32000 samples)
- Embedding dimension: vector of 96 features

II. Use cases

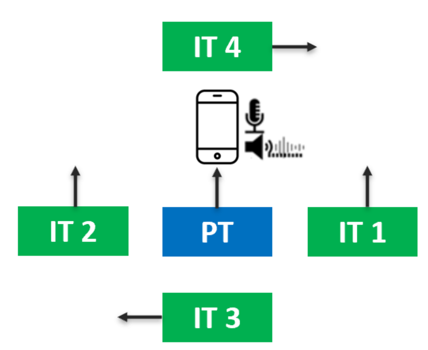


Challenges:
- Speaker ID network introduces the decision latency
- Robustness to false detections

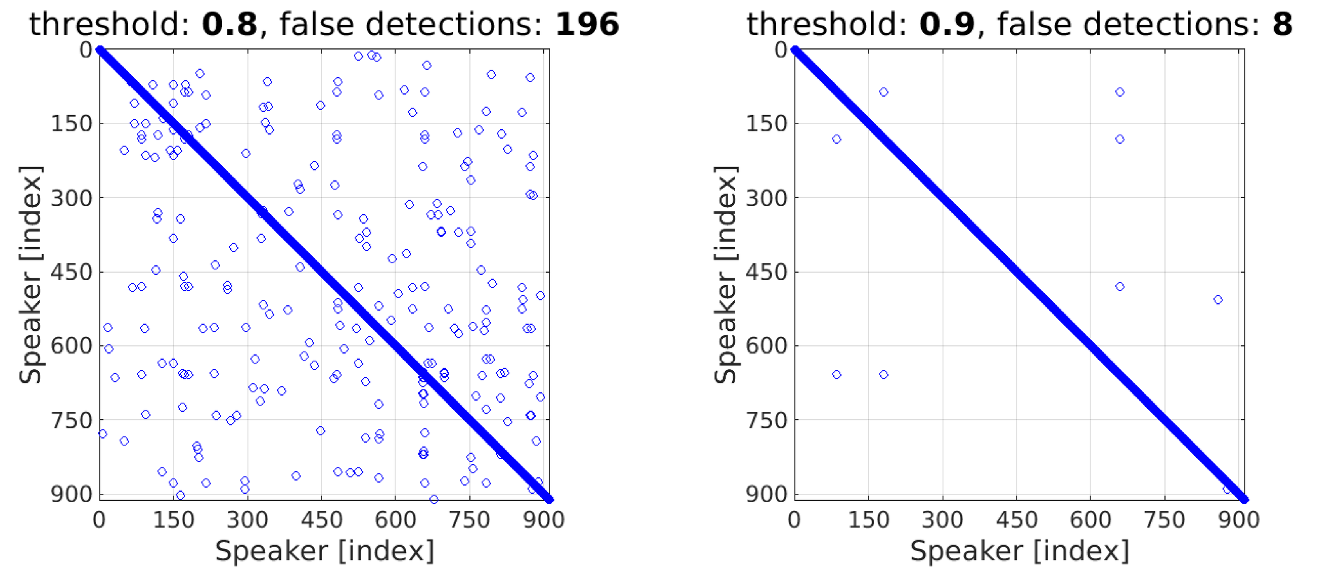
III. Example



PT – primary talker
IT – interfering talker
no overlap, ITs 1 m from PT

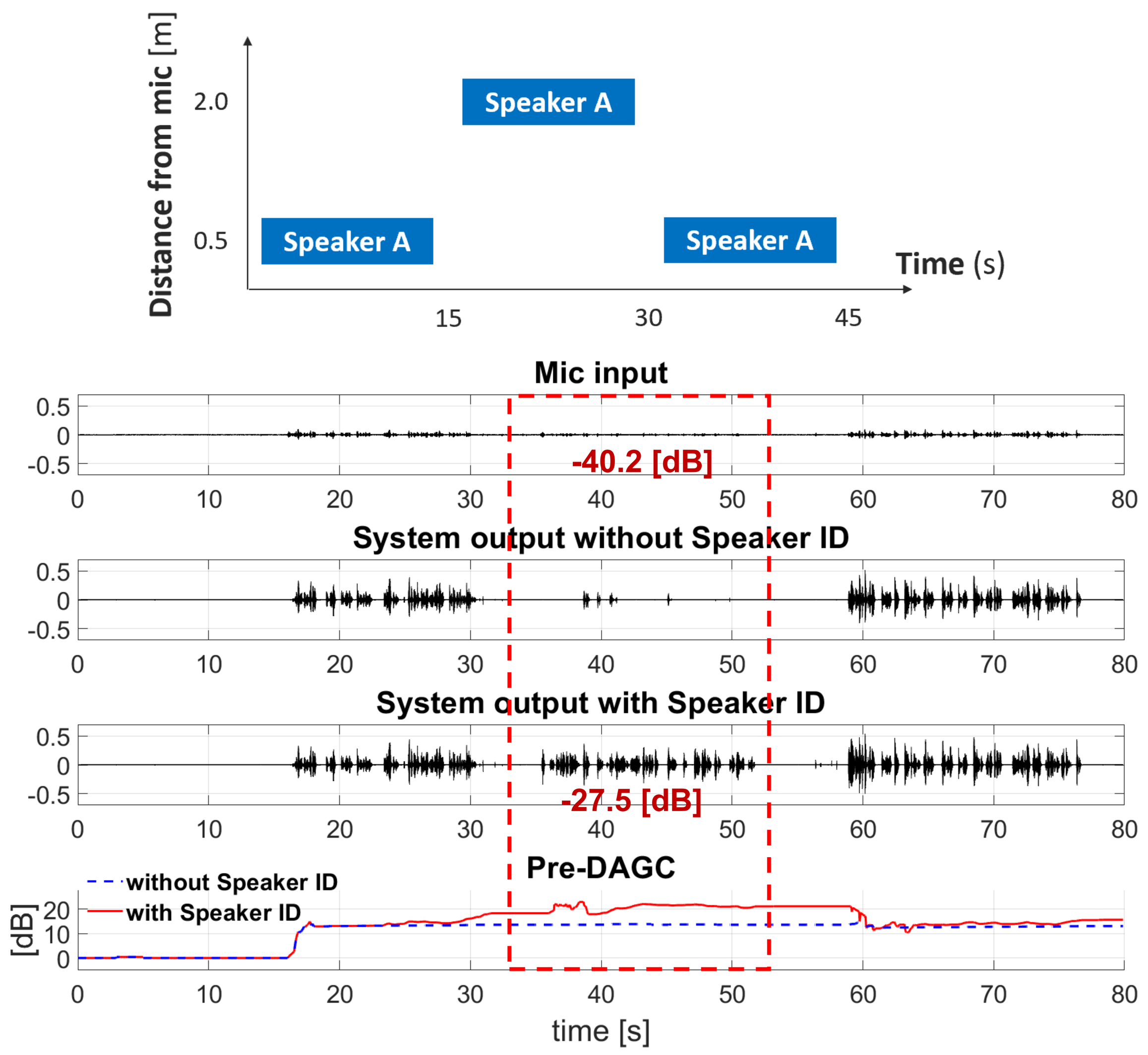


IV. Similarity matrix for clean voices

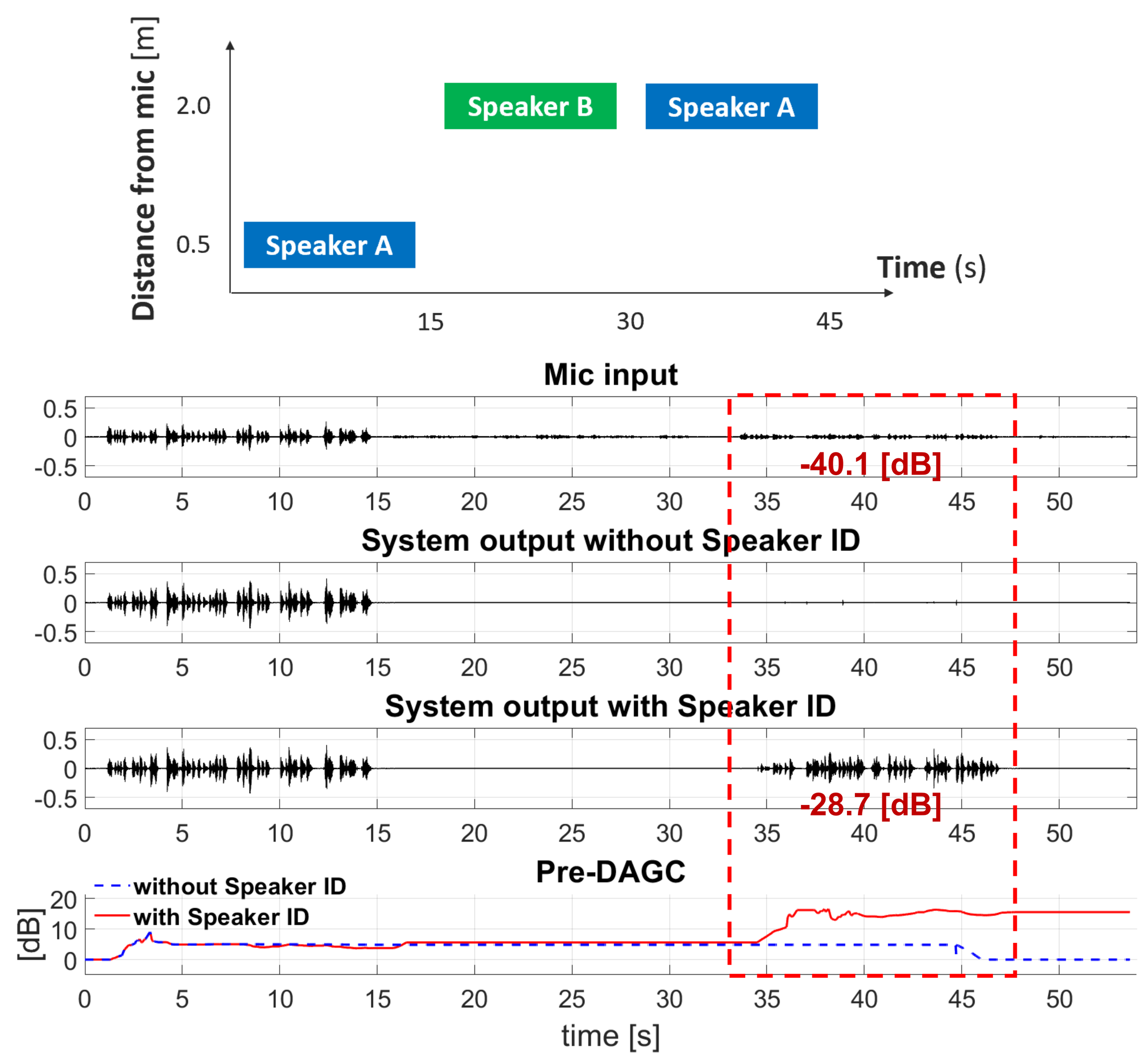


4. Demo scenarios

I. Case A

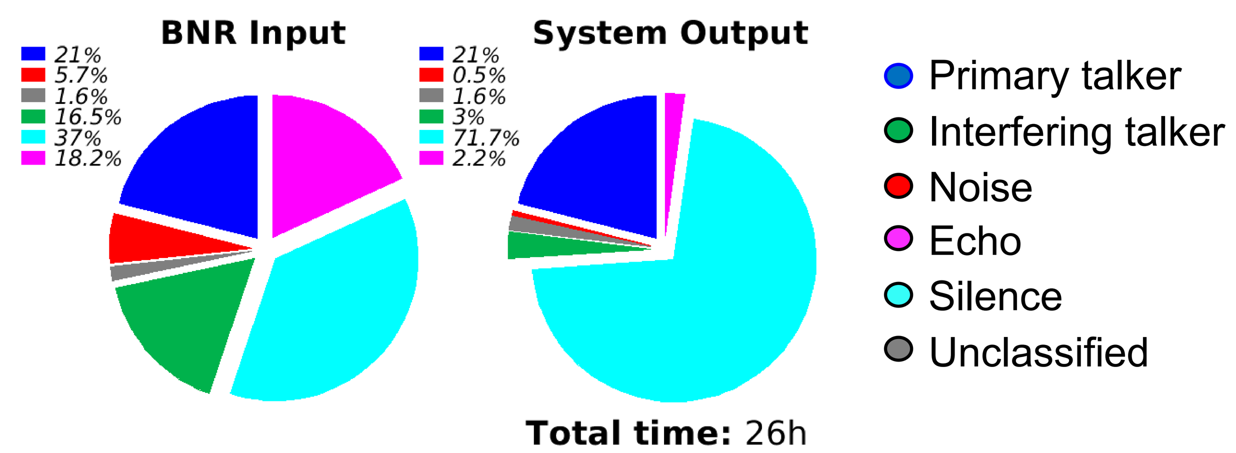


II. Case B

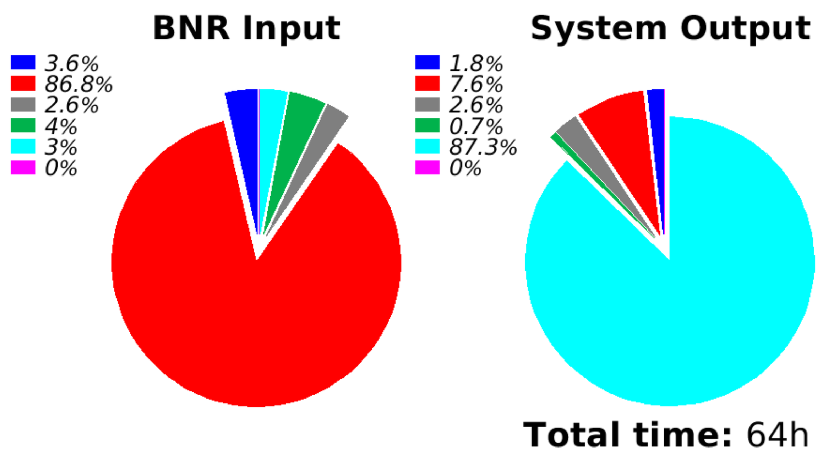


5. System results

I. Live recordings



II. 24 noise categories



- keyboard striking
- vacuum cleaner
- kitchen noises
- babble noise
- dog barking
- baby crying
- car noise
- music

6. Future work

Optimize for my Voice with primary talker identification based on audio and video.