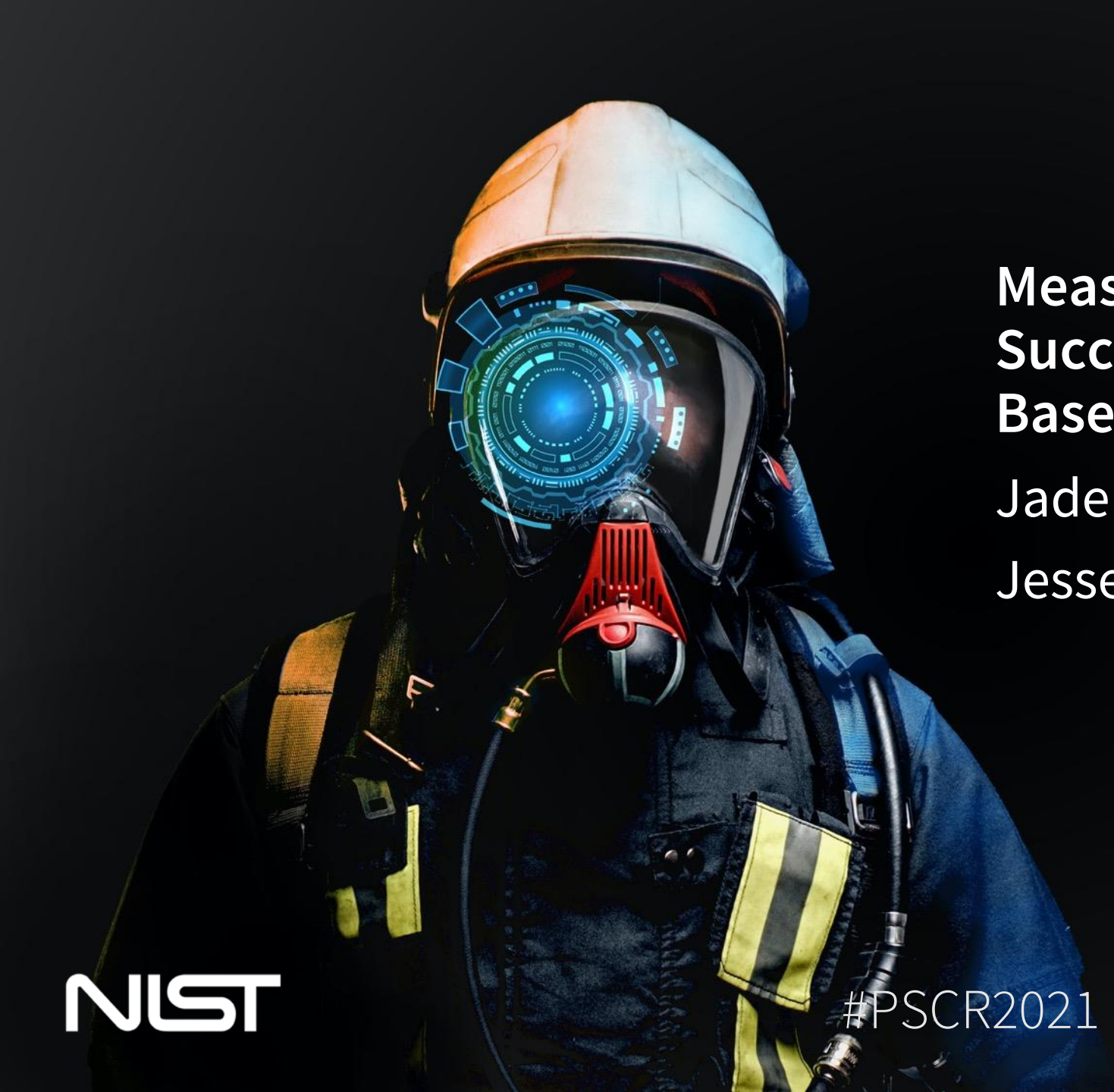

PSCR 2021
THE DIGITAL EXPERIENCE
#PSCR2021 • PSCR.GOV
NIST
PSCR



Measuring the Probability of Successful Delivery: a QoE Based Approach

Jaden Pieper PSCR MCV

Jesse Frey PSCR MCV

NIST

#PSCR2021

PSCR

DISCLAIMER

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately.

Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

*** Please note, unless mentioned in reference to a NIST Publication, all information and data presented is preliminary/in-progress and subject to change**

Overview

- Quality of Experience Based Measurements
- Probability of Successful Delivery Definition
- Intelligibility
- Measurement Technique
- Validation
- Future Work

Quality Of Experience (QoE) Based Measurements

- QoE KPIs for Mission Critical Voice (MCV)
  - Mouth-to-Ear Latency
  - Access Time
  - Voice Quality/Intelligibility
  - Probability of Successful Delivery (PSuD)

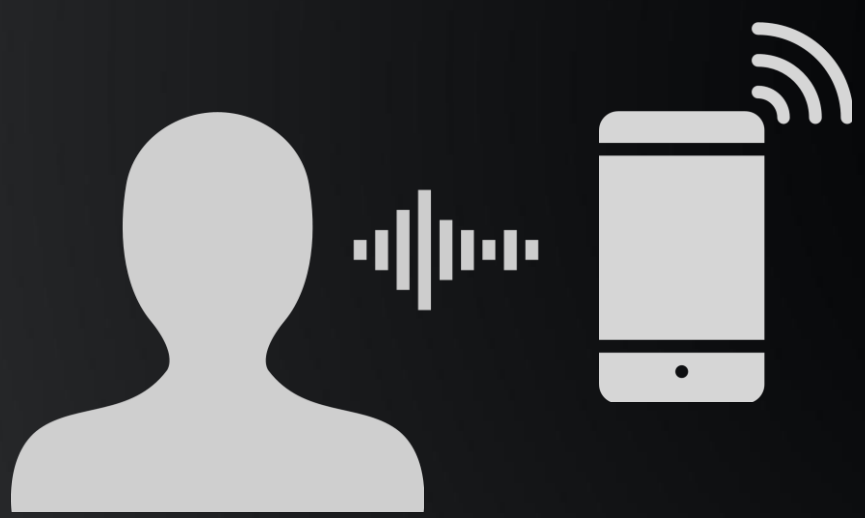
The User Experience: PTT Communications

- Press the Push To Talk (PTT) button and speak into a device
- Listening to speech output from a device
- It is all about speech

- Goal - Create measurement systems that are:
  - Based upon the user experience -- speech
  - Comparable and fair across technologies
- This is not:
  - Analyzing internal system design/construction

Technology Agnostic Measurements

Communications System



KPIs:

- Mouth-to-ear Latency
- Access Time
- Intelligibility
- Probability of Successful Delivery

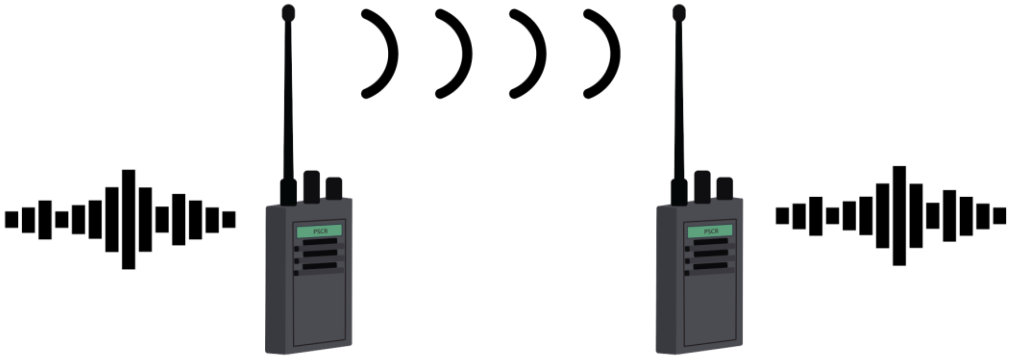
Probability of Successful Delivery

- Probability of Successful Delivery (PSuD)
  - **Definition:** The probability of successful delivery, $P_S(T)$, is the probability of successfully transmitting and receiving a message of length T.
- Encompasses probability of access and retention
  - Hard to separate
  - Roll into one
- Length of your message matters
- What is success?
  - Achieves an intelligibility threshold

Probability of Successful Delivery

Send the same message 5 times

Radio



Full message intelligible?

Examples of received messages:

Radio | Ra...o | ...dio | Radio | Radi...

Probability of Successful Delivery of "Radio" = 2/5 = 40%

Intelligibility

- Modified Rhyme Test (MRT)
  - Batches of six words
    - *went, sent, bent, dent, tent, rent*
    - Words within a batch rhyme
    - 50 batches total, 300 MRT keywords
  - MRT Trial
    - Carrier phrase + word
    - “Please select the word *went*”
    - Success (identified) or Failure (mis-identified)
- ABC-MRT16
  - Objective algorithm to estimate MRT scores
  - Get results “on demand”
- For more information, view the on-demand session:
  - Designing Remote Listening Experiments for the Partially Muted Word Impairment

Measurement Strategy

- Send strings of MRT keywords through a system
  - From ABC-MRT16 database[1]
  - No carrier phrase, keywords only
- Measure intelligibility of each word
- Determine if message was a success or not
- Do this over lots of trials

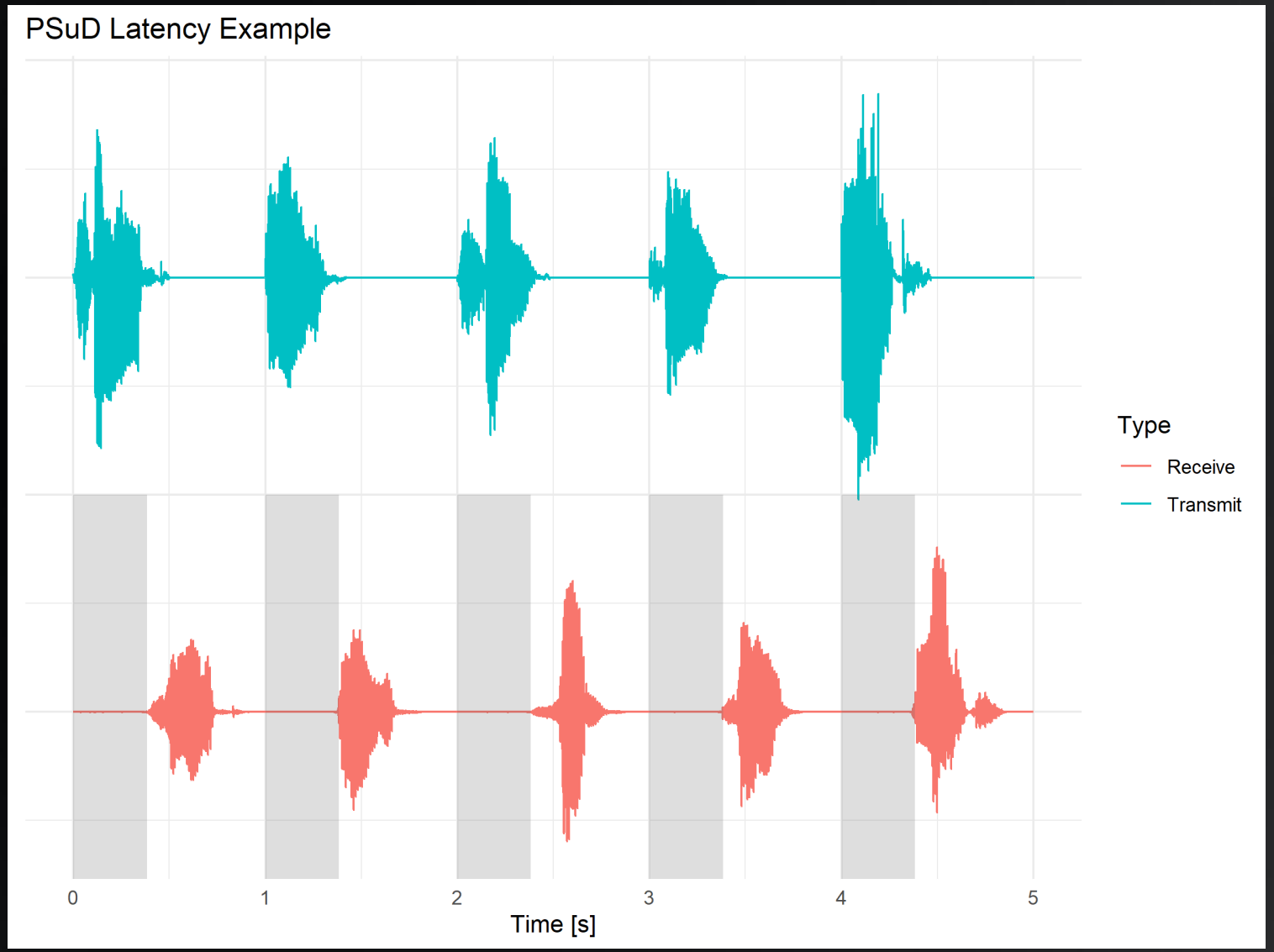
1: Voran SD (2017) A multiple bandwidth objective speech intelligibility estimator based on articulation index band correlations and attention. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 5100–5104.doi: 10.1109/ICASSP.2017.7953128

Audio Clip Structure

- 10 second clips, one MRT keyword every second
- Highly structured, know where words should be
- Good neighbors
  - No keywords near keywords from shared batches
  - “look” is not next to “hook”
  - Prevents poor alignment causing estimate errors
- MRT Coverage
  - Broad sense of system performance
  - 10 words per clip
  - 30 clips is full coverage for a talker
  - 4 orderings to minimize additional vocoder effects
  - 120 clips per talker, 480 total (4 talkers)

Latency Estimation

- Need to know what word is where
- Challenging when lots of audio is missed
- Clip structure alleviates this
- Can eyeball latency



Classifying Success

- Intelligibility thresholds
- What about context?
  - Normal speech tends to have it
  - Special cases do not
    - Phone numbers, license plates
    - Addresses if you do not know the area
- No context is worst case scenario for PSuD
  - Cannot miss a single word
- MRT strings lack context
  - Simulates special no-context cases perfectly!
  - How do we simulate context?
  - PSuD needs to be more robust than the worst-case scenario

Simulating Context

- Every Word Critical: EWC
  - Worst case
  - No context
- Average Message Intelligibility: AMI
  - Best case
  - Context distributed uniformly through a message
- Autoregressive Filter: AR(1)
  - Somewhere in the middle
  - Filter weight gives us a dial to play with
  - Simulates message context memory

Every Word Critical (EWC)

- Success – every word of the message achieves the threshold intelligibility
- No context in the message
- Worst case scenario
  - Lose one word, the rest of the message fails

Example Data - P25 Direct
Threshold = 0.7

Word	W0	W1	W2	W3	W4	W5	W6	W7	W8	W9
Word Intelligibility	1	0.4375	0.75	0.875	1	1	1	1	1	1

Average Message Intelligibility (AMI)

- Success – average intelligibility over the whole message exceeds a certain threshold
- $\text{Output}(k) = \frac{1}{k+1} * \sum_{j=0}^{k} \text{Input}(j)$
- Context is uniformly distributed throughout the message
- Best case scenario: lose one word, every other word provides context

Example Data - P25 Direct
Threshold = 0.7

Word	W0	W1	W2	W3	W4	W5	W6	W7	W8	W9
Word Intelligibility	1	0.4375	0.75	0.875	1	1	1	1	1	1
Filtered Intelligibility	1	0.7188	0.7292	0.7656	0.8125	0.8438	0.8661	0.8828	0.8958	0.9063

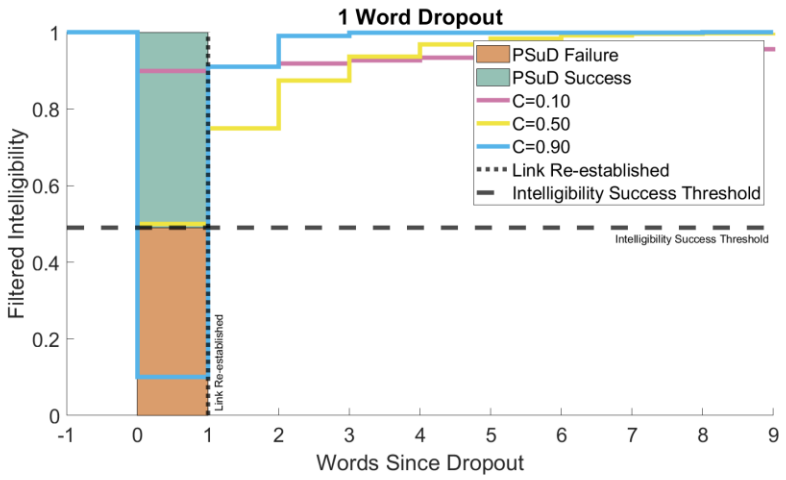
Autoregressive Filter - AR(1)

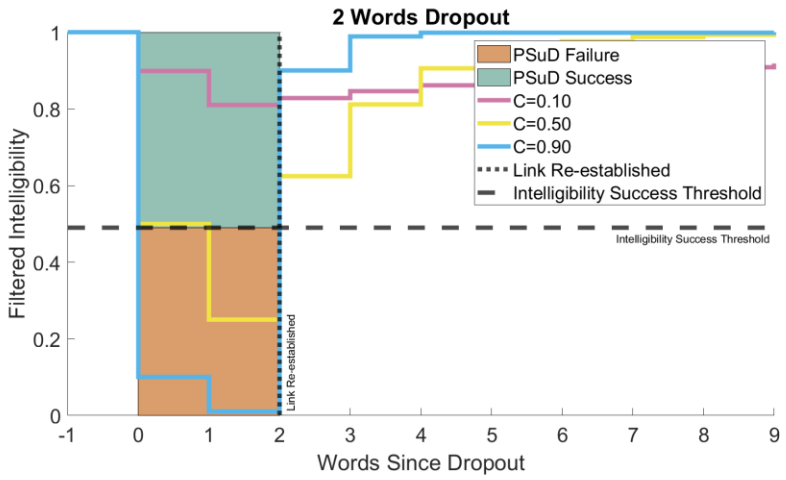
- Success – Filtered intelligibility for each word of the message exceeds a certain threshold
- $\text{Output}(k) = c * \text{Input}(k) + (1 - c) * \text{Output}(k - 1)$
- Context decays as you move away from a previous word
- Middle ground
  - Filter parameter controls decay rate

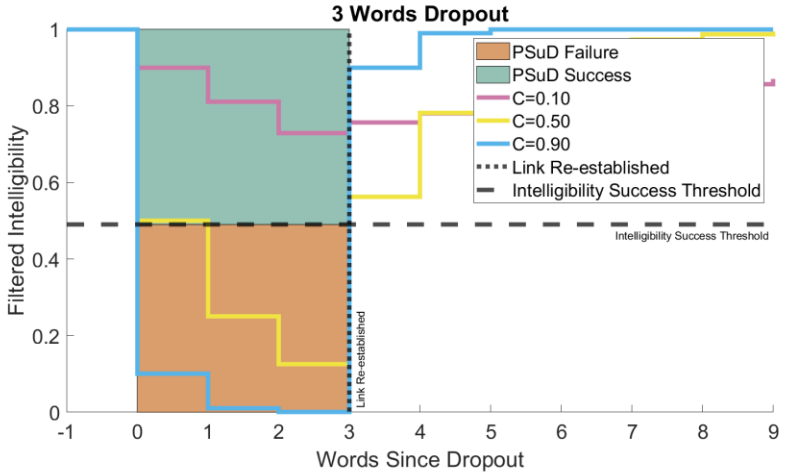
Example Data - P25 Direct
Threshold = 0.7
$c = 0.5$

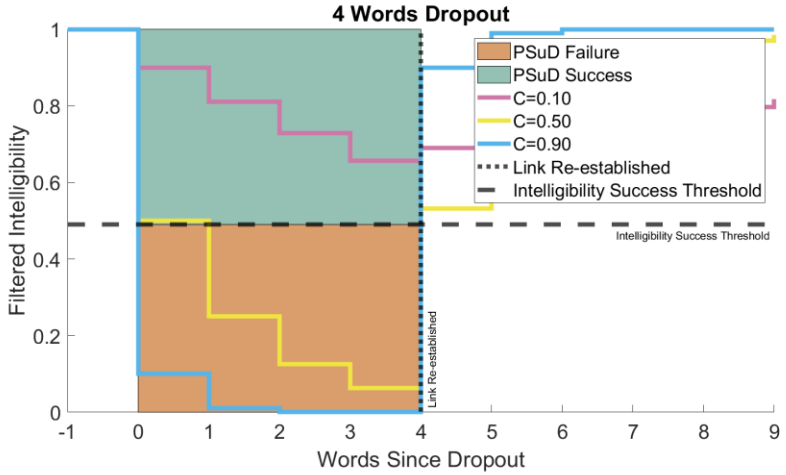
Word	W0	W1	W2	W3	W4	W5	W6	W7	W8	W9
Word Intelligibility	1	0.4375	0.75	0.875	1	1	1	1	1	1
Filtered Intelligibility	1	0.7188	0.7344	0.8047	0.9023	0.9512	0.9756	0.9878	0.9939	0.9969

Autoregressive Filter – AR(1)

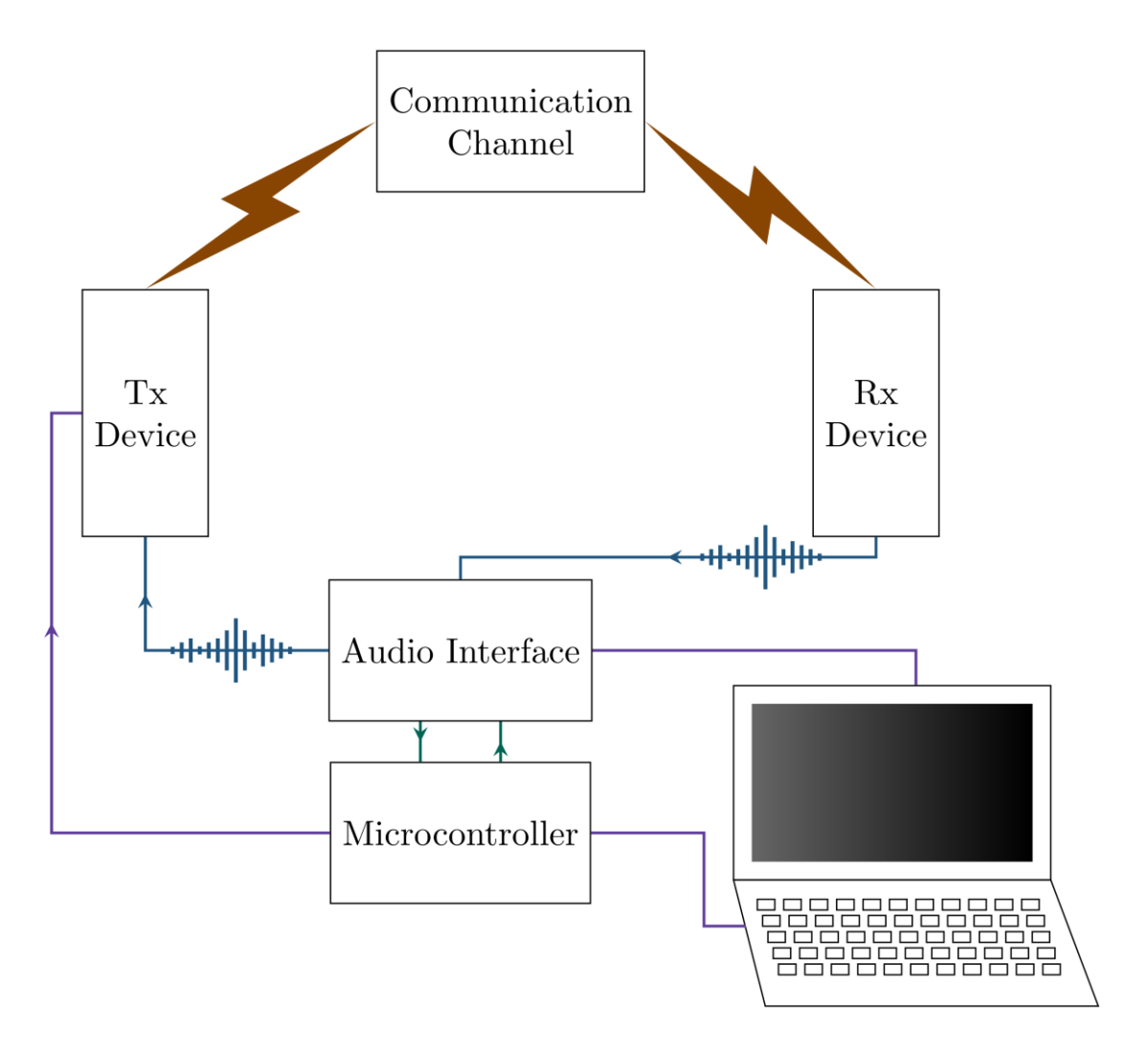








System Diagram



Measurement Software

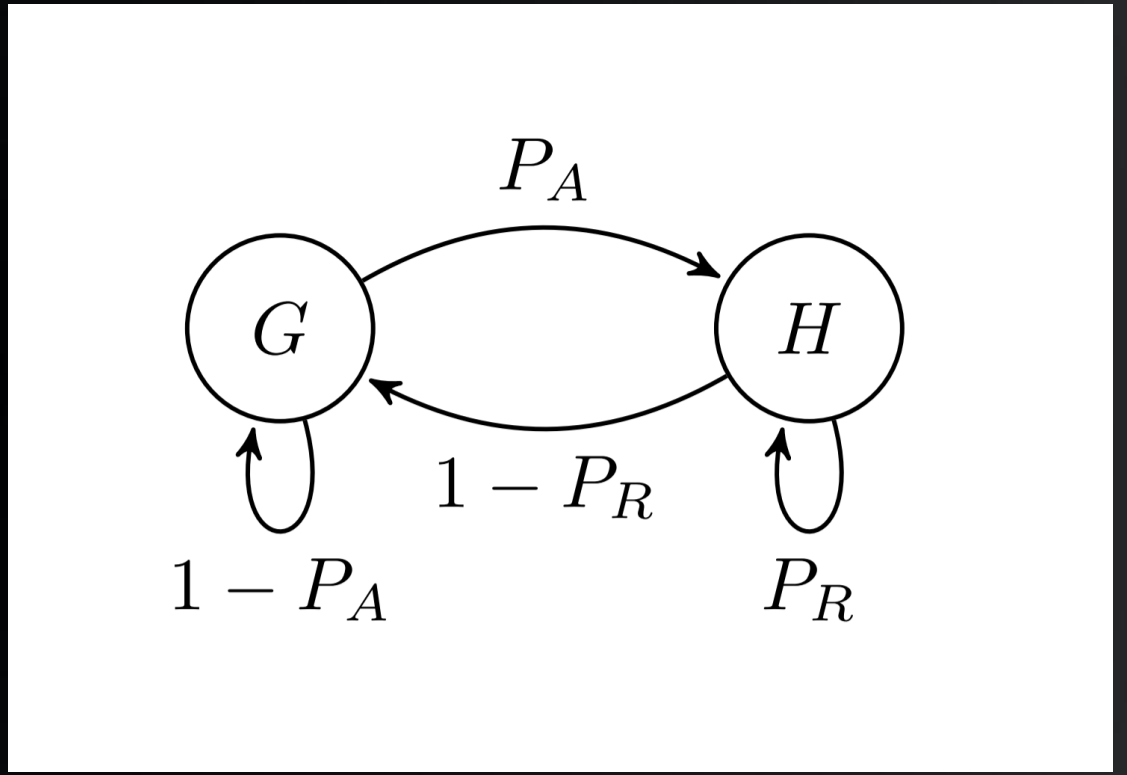
- Developed in Python
  - MCV shared code modules
- Improved simulation features
- For more information, please view the on-demand session:
  - QoE Software and Hardware Packaging

Validation

- Need to know measurement is accurate
- Simulated measurements
  - Software implementations
    - Radio interface
    - Audio interface
- Control impairments directly
- Verify results match expectation

Validation

- Model
  - G - communication links not established
  - H - communication link established
  - P_A - probability of establishing communication link
  - P_R - probability of retaining communication link
- Markov Model
  - Every word critical (EWC) case
  - PSuD of T second long method, state changes every c seconds.
  - $P_S(T) = \frac{P_A}{1+P_A+P_R} * P_R^{\left(\frac{T}{c}\right)-1}$



Future Work

- Continue developing context methods
- Validating measurement strategy is successful on more technologies
- Radio frequency (RF) channel impairments
- Two-location measurements

Related PSCR 2021 On-Demand Sessions

- Mission Critical Voice Quality of Experience Measurement Methods Overview
- Introducing a Start of Word Correction for Access Delay Measurements
- Designing Remote Listening Experiments for the Partially Muted Word Impairment
- Optimal Transmit Volume Conditions for MCV QoE Measurement Systems
- QoE Software and Hardware Packaging
- QUARC: Quality Under Adjustable Realistic Conditions for Communication Systems
- Lab from Home: Distributed QoE Testing for Mission Critical Voice

Questions

- Looking forward to answering questions at our live Q&A session!

THANK YOU

#PSCR2021 • PSCR.GOV