Speech Analytic Technologies Performance Evaluation Project (OpenSAT)

Fred Byers, NIST/ITL

July 9, 2019

#PSCR2019

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**

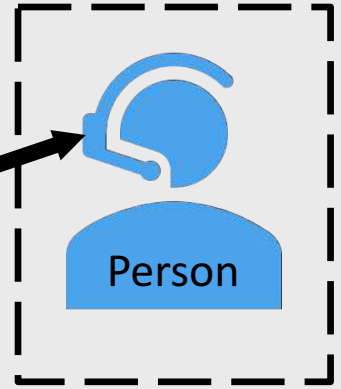
Previous vs Current Work

Previous work

Purpose was to quantify speech codec intelligibility for humans.

Speech recorded then noise added

+ Codec effects on speech



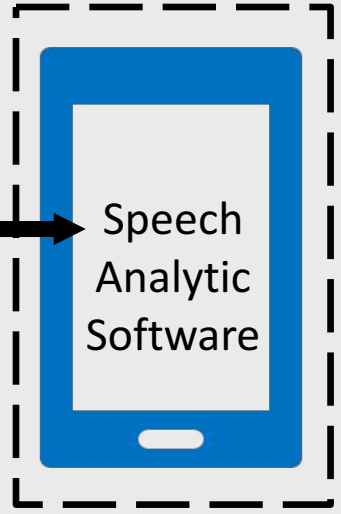
Human intelligibility

Current work

Purpose is to evaluate speech analytic technology capability.

Speech recorded in presence of noise

+ Noise effects on speech



Technology capability

Agenda

- Project Overview
- Challenges
- NIST Research Model
- OpenSAT
- Next Steps

Sponsored by Department of Homeland Security Science and Technology Directorate (DHS S&T)

The Problem

Performance of assistive technologies can degrade significantly in first responder scenarios.

There are currently no relevant data sets to evaluate speech analytic technologies for first responder applications.

Speech Analytic Challenges
In the Public Safety Domain

Background Noises and Altered speaking

- Existing data sets lack the impact on the speaker while in first responder related background noises.

Why?

Environmental Noise Conditions

- **Multiple Types of Sounds**
- **Varying Volume Levels**
- **Overlapping Background Speaking**
- **Crowd Noise**

Speech-Effect Conditions

- **Lombard Effect**
- **Stress**
- **Acoustic Capture**
- **Transmission Effects**

Evaluating Speech Analytic Technology Performance

NIST began in 1987

The evaluation-driven research paradigm has been successful for many speech analytics.

Level Playing Field

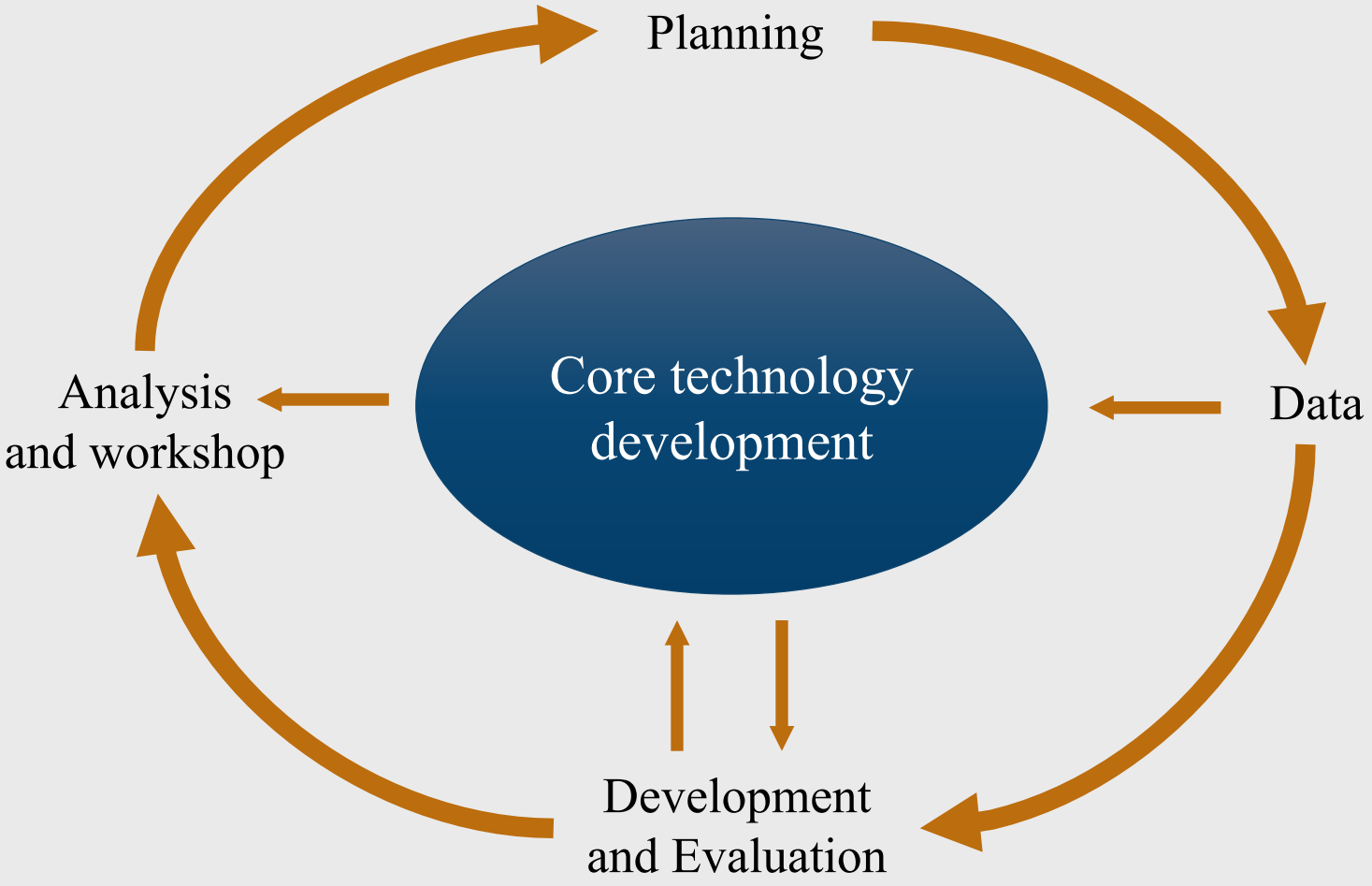
Alternative speech analytic algorithms use the same data and are evaluated by the same metrics.

Project Goal

Implement a relevant methodology that helps advance speech analytic technologies for public safety applications.

Evaluation-Driven Research Model

NIST provides the infrastructure for the outer ring.



Open Speech Analytic Technologies (OpenSAT)
Evaluation Series

OpenSAT19 is the first evaluation in the OpenSAT series.

Open	anyone can participate
SAT	Speech Analytic Technologies
Evaluation	an event covering several months
Series	event held annually

OpenSAT Highlights

Data

1. Simulated Public Safety Communications,
2. Video Annotation for Speech Technologies (VAST),
3. Babel

Tasks

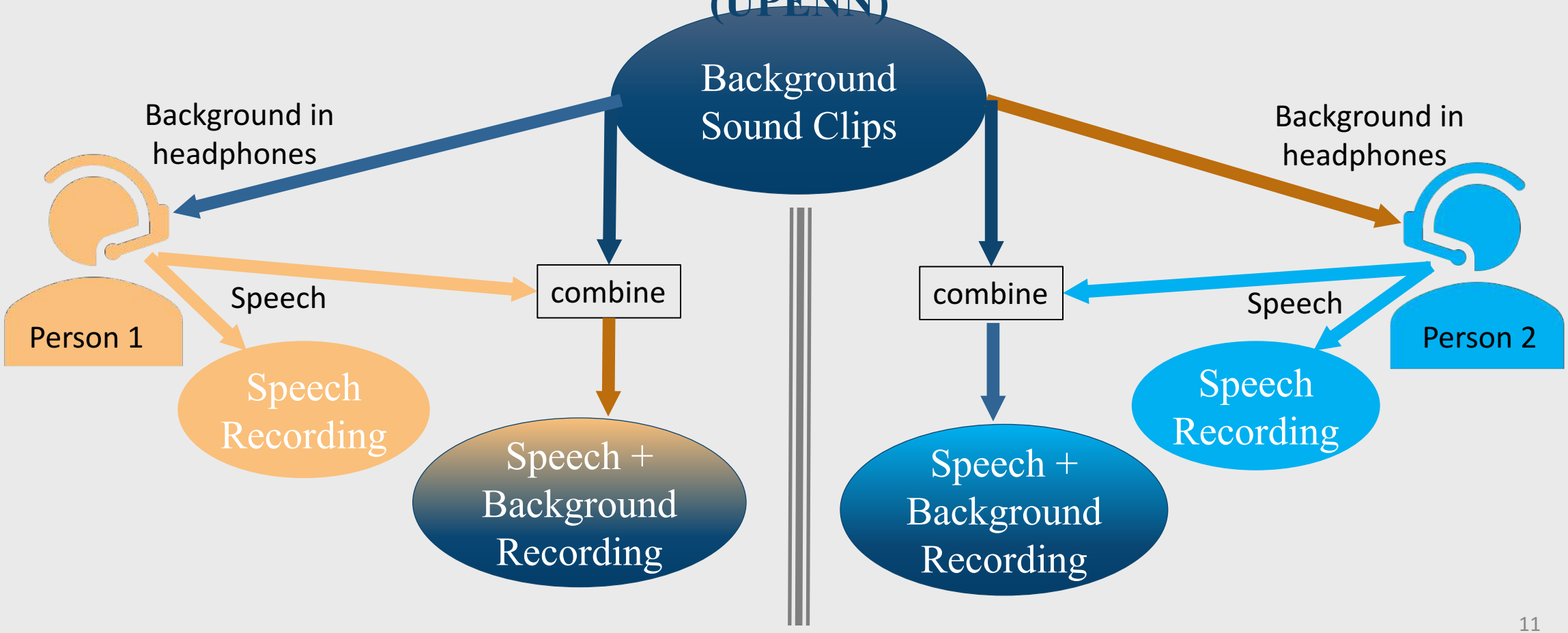
1. Speech Activity Detection,
2. Automatic Speech Recognition,
3. Keyword Search

OpenSAT Pilot - 2017

Results affirmed the enormous challenge for speech analytic technologies in real world first responder operational scenarios.

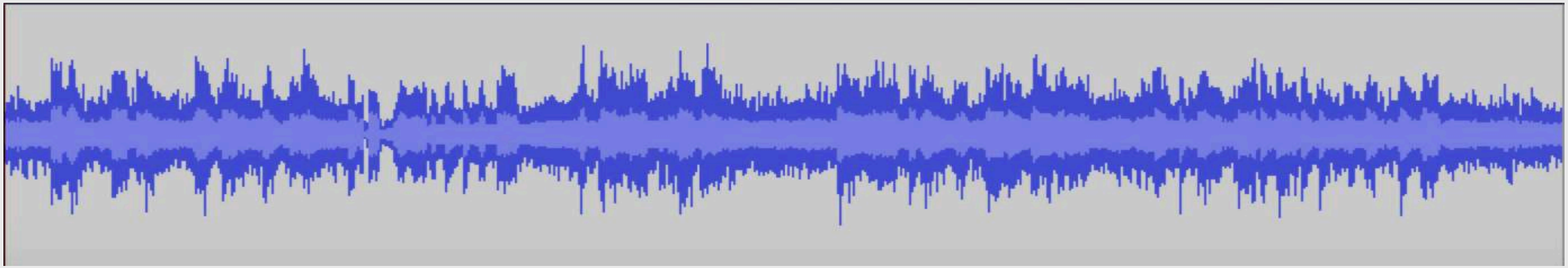
Simulated First Responder Communications Game Recording

Linguistic Data Consortium (LDC) at University of Pennsylvania (UPENN)

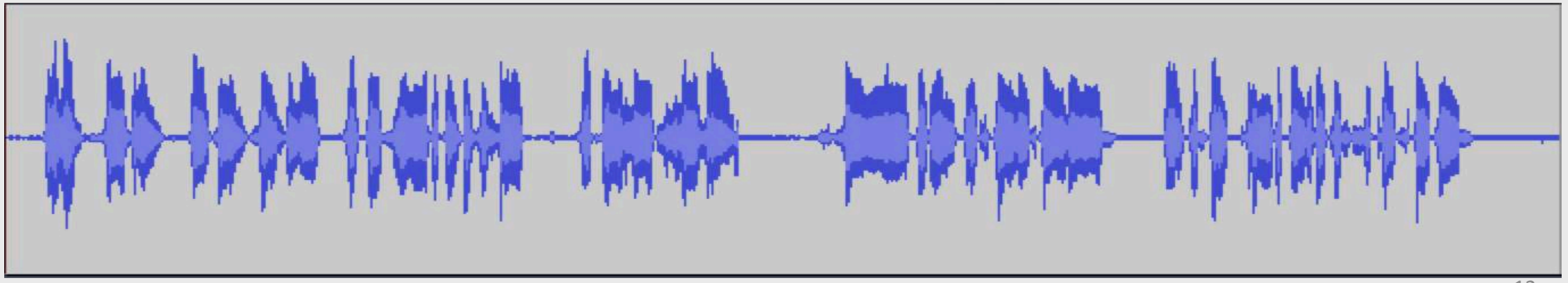


Background Noise and Speech
(Same Speech in Both Examples)

Background + Speech



Speech only



OpenSAT Benefits

- Public safety relevant data for development and testing
- Assessment forum for underlying technologies in public safety applications
- Understanding for developers in public safety domain challenges
- Technology advancement
- Baseline and tracking
- Agencies/organizations can evaluate potential solutions

Next Steps

After the OpenSAT19 Evaluation

Next year

- OpenSAT20
- Using new evaluation data from LDC's data set
- Track progress

Future

- Obtain recordings of real world first responder operational communications
- Have these recordings annotated and converted to a usable format
- Create new data sets from these recordings with best quality control available
- Implement the new data sets into future OpenSAT Evaluations

Contact Us

Fred Byers – NIST/ITL/ Information Access Division: Multimodal Information Group

frederick.byers@nist.gov

https://wwww.nist.gov/itl/iad/mig/opensat

https://sat.nist.gov

THANK YOU

#PSCR2019

Come back for the

Next Session

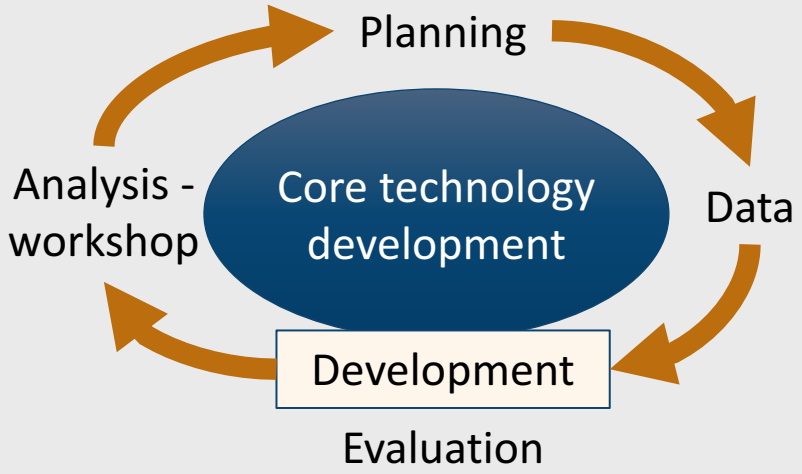
1:35 PM

Backup Slides

OpenSAT Evaluation Protocols

Development Phase

Data	Developer use	Audio size	Annotation style	Annotation given to developers?
Training	Build system	100+ hours	Light	Yes
Development	Self test	5 hours	High quality	Yes

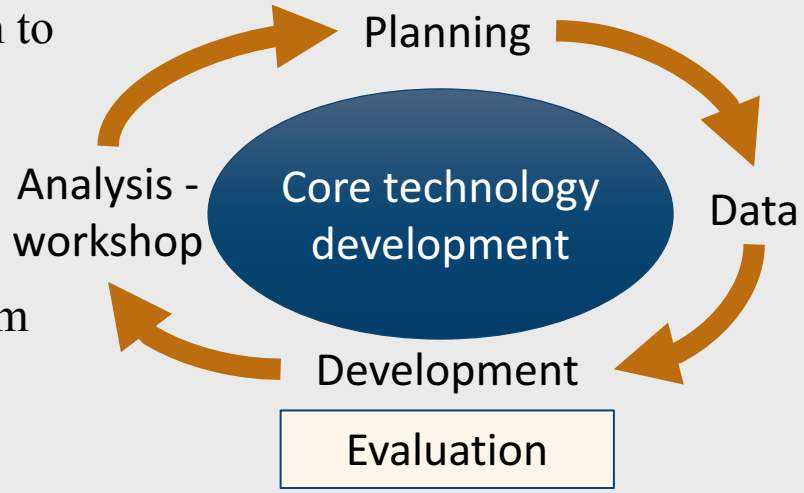


OpenSAT Evaluation Protocols

Evaluation Phase

Data	Developer use	Audio size	Annotation style	Annotation given to developers?
Evaluation	The “test”	5 hours	High quality	No

1. Developers apply their system to the evaluation data.
2. Developers upload their system output through the NIST web site.



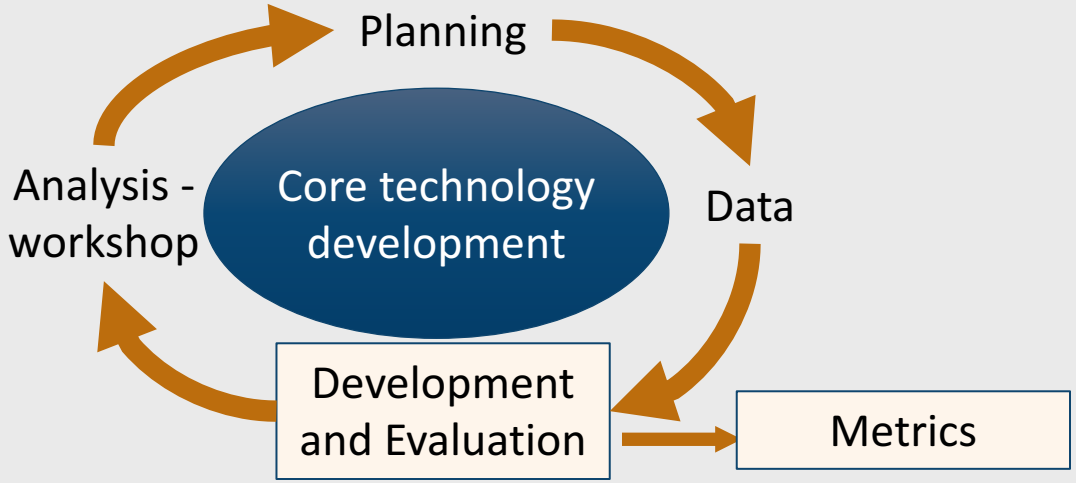
3. Scoring server measures performance and returns a score
4. Scores are ranked relative to the pool of uploads

Metrics

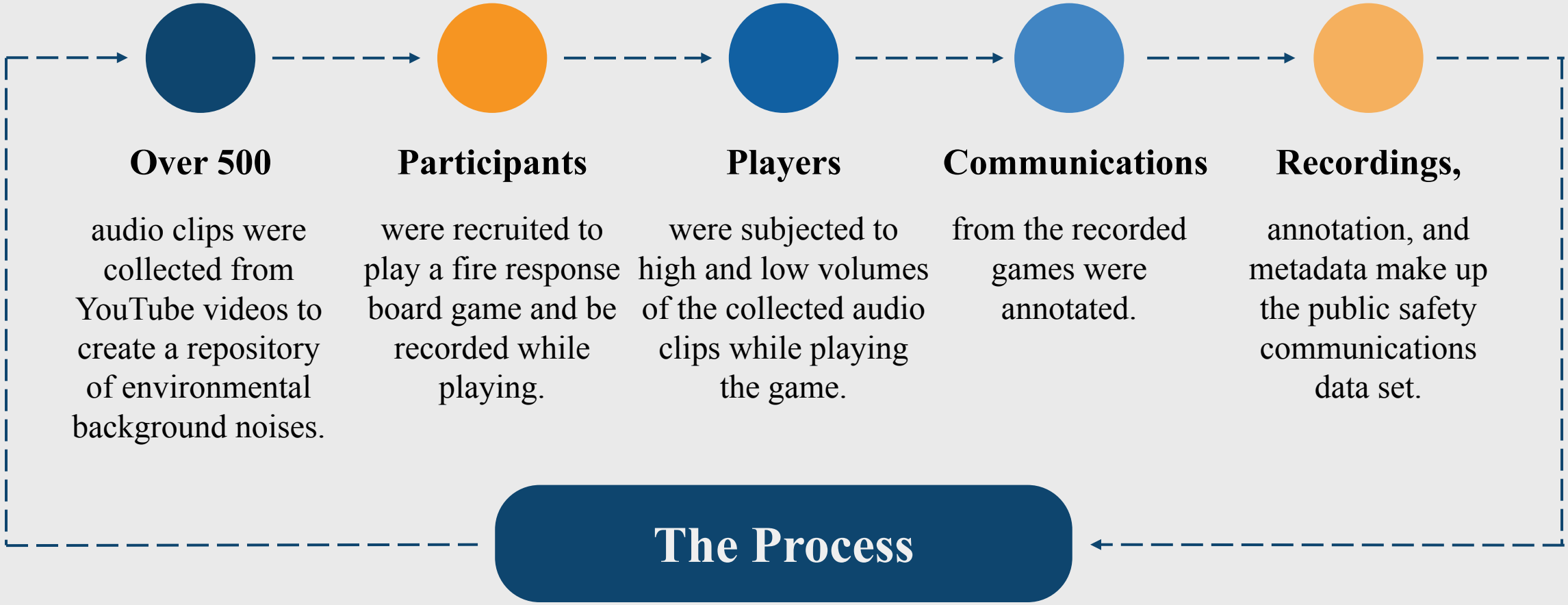
Task	Metric
Speech Activity Detection	Detection Cost Function value (DCF)
Automatic Speech Recognition	Word Error Rate (WER)
Keyword Search	Actual Term Weighted Value (ATWV)

In Summary:

All three metric values are a function of the number of errors produced during the test.



Public Safety Communications Data Collection



Audio Collection by the Linguistics Data Consortium (LDC) at UPENN

- Participants played in pairs.
- Multiple sessions/games per player.
- Background noises were fed into participants' headsets.
- Volume varied from below to above the threshold that induces the Lombard Effect.
- Time constraints were imposed to create a sense of urgency.
- Participants could not see each other while playing the game.

Flash Point Fire Rescue Board Game

Game-Playing for Speech

30

MINUTES

Two 30-minute games were played per session to limit participants' vocal straining.

100+

PARTICIPANTS

Over 100 participants were recruited to include a diversity of speech characteristics.

50%

AGES 30-50

The majority of participants were 30-50 years old, 25% of participants were under 30 years old and 15% were over 50 years old.

50%

LOCATION

Fifty percent of the participants were from the Philadelphia area. Other participants stemmed from CA, NJ, MD, MT, TX, MN, MA.

Pilot Evaluation Results

Data set consisted of real world operational dispatches.

Plots show performance levels from multiple systems.

High error rate for speech-to-text transcription.

High level of missed speech in speech detection.