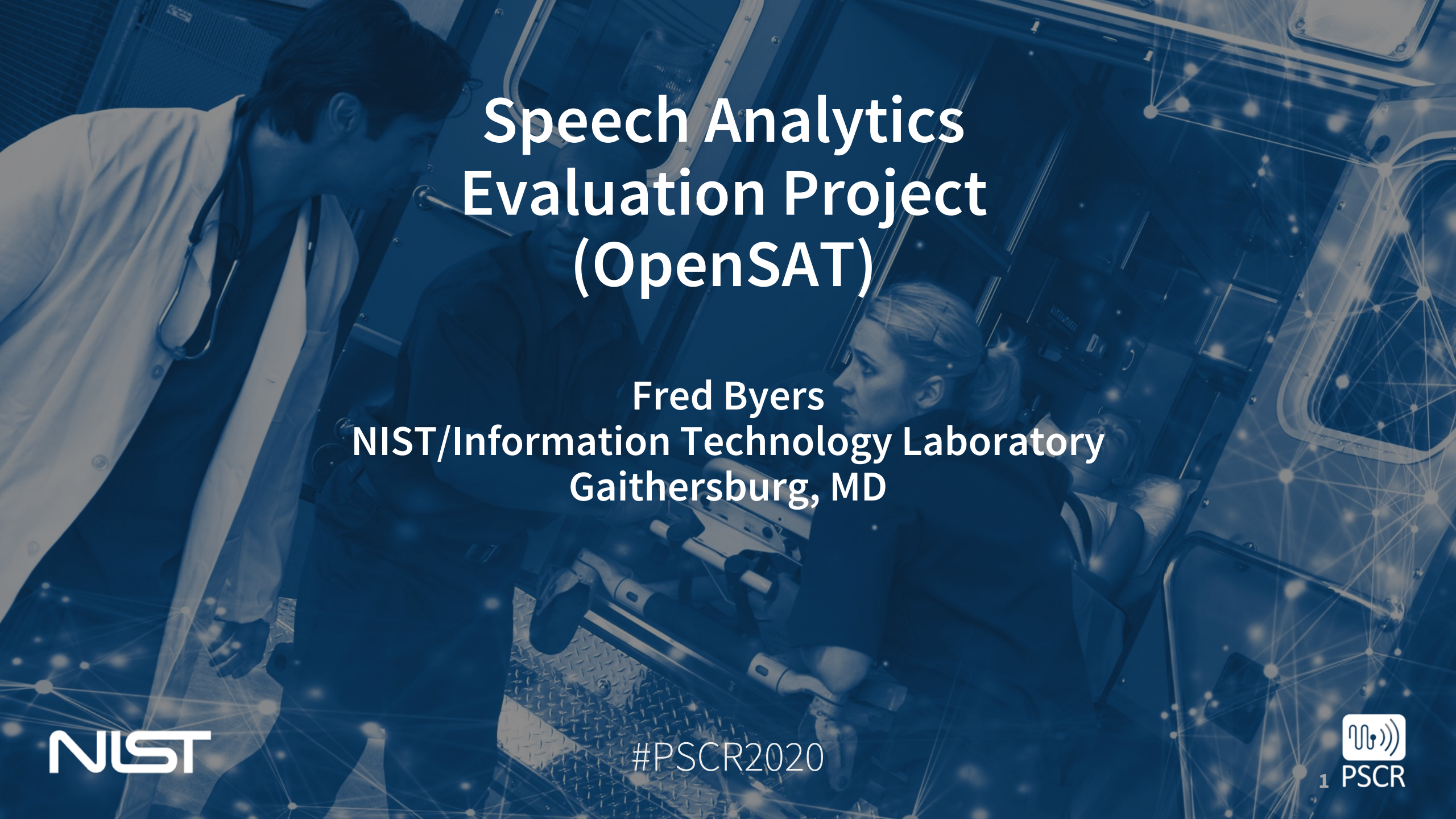


Speech Analytics Evaluation Project (OpenSAT)

Fred Byers
NIST/Information Technology Laboratory
Gaithersburg, MD

NIST

#PSCR2020

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**

Agenda

- Section 1
  - OpenSAT overview
  - Analytics (Tasks)
  - Data
  - Evaluation process
- Section 2
  - Equations and 2019 Test Results
  - Moving forward

OpenSAT Evaluation Series

NIST Open Speech Analytics Evaluations

Open
Anyone can participate

SAT
Speech Analytic Technologies

Evaluation
A process lasting several months

Series
The Evaluation is reoccurring

OpenSAT Speech Analytics Testing

Goal

Advance the performance of evolving applications that rely on speech analytic technologies.

Purpose

To facilitate data-driven research to evaluate and improve the performance of speech analytics for public safety applications.

Speech Analytic Challenges
in Public Safety Communications

Acoustics

Loud background sounds

Varying volume levels

Multiple types of background sounds

Speakers in the background

Crowd noise

Type of mic

Speech

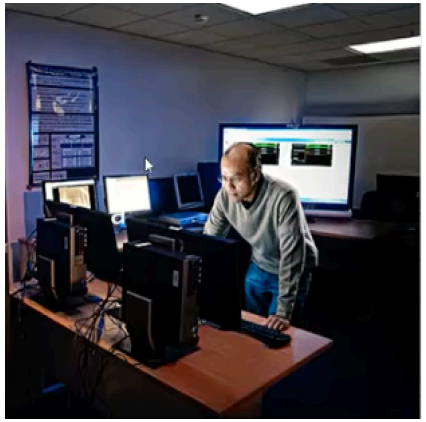
Increased vocal effort (Lombard effect)

Urgency

Stress

How OpenSAT Evaluations Improve Applications

Researchers participate in OpenSAT



Multiple research teams simultaneously and independently

- work with the same data and the same metrics
- focus on common speech analytic technologies (tasks)
- train and develop these tasks for specific data
- improve the core technologies that applications depend on

Tapping the researcher competitive spirit

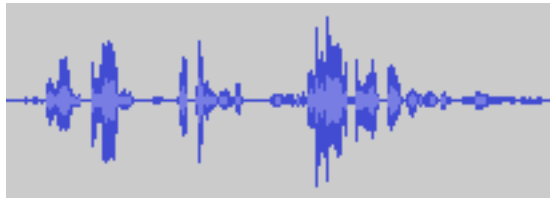
- NIST provides researchers the infrastructure, tools and testing model
- researchers compare system performance from the evaluation data
- state-of-the-art solutions are brought to the industry by researchers
- industry can compare their technology performance against others

The Core Technologies (Tasks) Evaluated

Currently, Three Tasks are Evaluated in OpenSAT

1

Speech Activity Detection



Automatically detect the beginning and ending of speaking segments by timestamps.

2

Automatic Speech Recognition

→ going inside now

Automatically transcribe verbatim all audible speech to readable text.

3

Keyword Search

***** alarm ***** ***
*** ****** *********
*** need backup ** ****

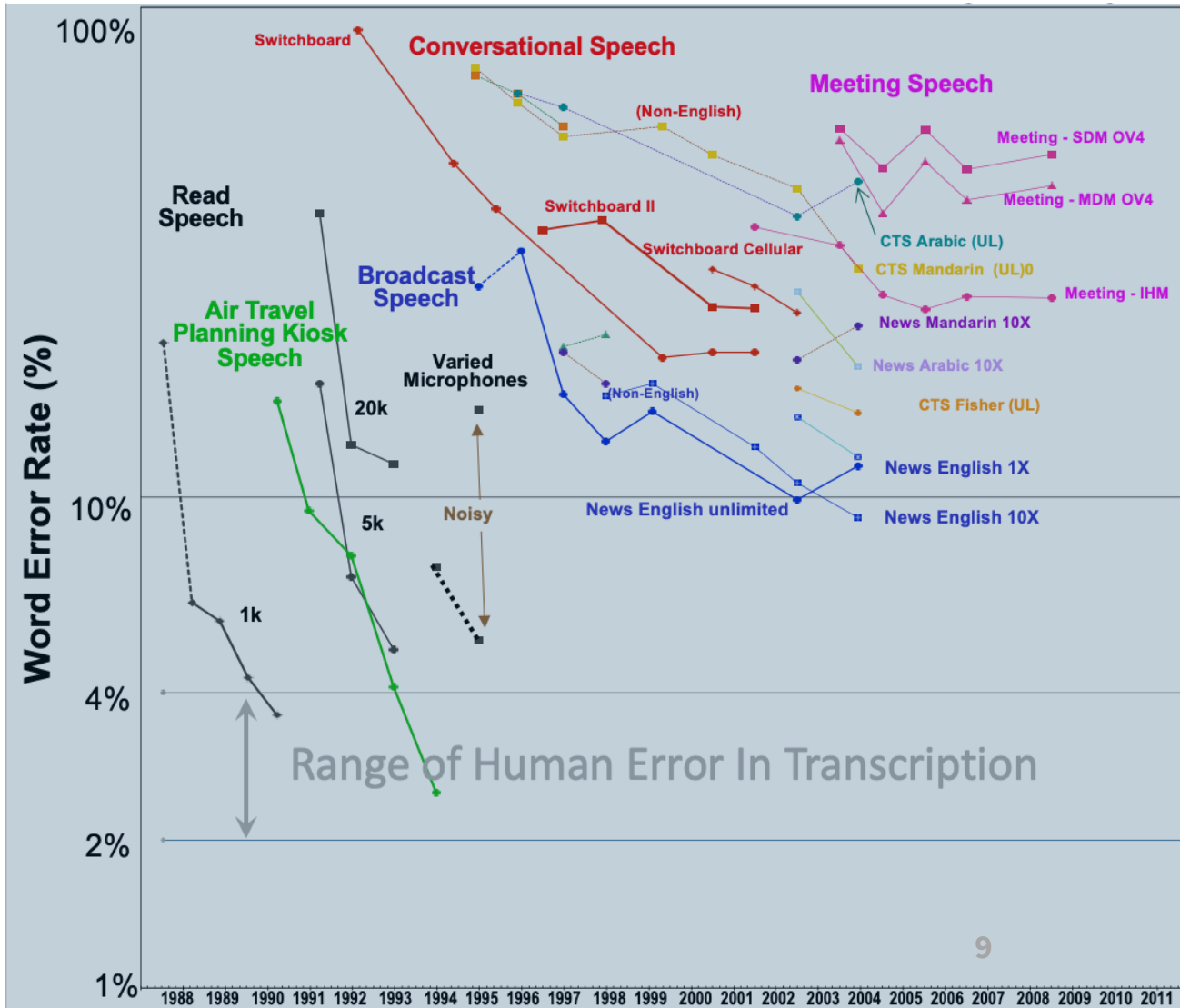
Automatically detect all spoken instances of a predetermined list of words and phrases.

To Evaluate You Need Relevant Data

Test data relevant to public safety

Relevant and high-quality data improves the evaluation of evolving technologies

- Hands-free voice interface
- Voice, video, GPS situational awareness system
- Automated transcription for reporting



The Linguistics Data Consortium (LDC) at the University of Pennsylvania (UPenn)

The Linguistic Data Consortium (LDC) recorded and transcribed dozens of fire rescue game-playing communications.

Audio recordings were collected in controlled conditions.

The collection of recordings and transcriptions were organized to create specialized data sets for distribution to researchers.

This collection was funded by DHS.

LDC Catalog	Corpus Name
	Researchers and NIST
LDC2019E37	SAFE-T Corpus Speech Recording Audio Training Data R1 V1.1
LDC2019E38	SAFE-T Speech Recording Development Audio and Transcripts
LDC2019E36	SAFE-T Speech Recording Training Data Transcripts V1.1
LDC2019E50	OpenSAT19 Public Safety Communications Simulation Evaluation Data V1.1
LDC2019E53	OpenSAT19 Public Safety Communications Simulation Development Data V1.1
	NIST only
LDC2019R03	SAFE-T Corpus - Speech Recording Metadata for Dev Selection
LDC2019R10	SAFE-T Corpus Speech Recording Audio R1 V1.1
LDC2019R09	SAFE-T Corpus Speech Recording Metadata for Eval Selection
LDC2019R17	SAFE-T Speech Recording Evaluation Data Transcripts V1.1
LDC2019R19	SAFE-T Speech Recording Evaluation Audio for NIST V1.2
LDC2019R22	OpenSAT19 PSC Simulation Evaluation Data for NIST Preview
LDC2019R25	SAFE-T Corpus Speech Recording Audio R2 V2.0
LDC2019R31	SAFE-T Speech Recording Development Evaluation Superset Audio
LDC2019R32	SAFE-T Speech Recording Training Data Transcripts R2
LDC2019R33	SAFE-T Speech Recording Development Evaluation Superset Transcripts
LDC2019R34	SAFE-T Corpus Speech Recording Audio Training Data R2

Audio Collection Process

Audio Collected

Communications are recorded during the game.

Background noises are fed through headphones in various noise types and varying volume levels.

Participants were recruited to play a collaborative board game to rescue victims from a burning fire.

First responder background noises are fed through headphones of participants while playing the game.

The background noises and sense of urgency elicit varying levels of vocal effort intended to simulate real-world operational speech.

Selected audio files were transcribed to create development and evaluation data sets.

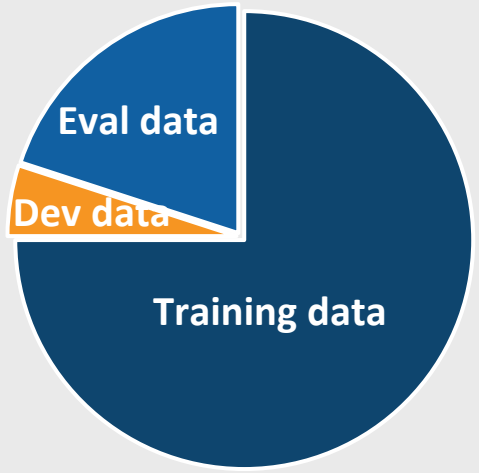
Data Sets Created

Audio Collected → **Data Sets Created**

Communications are recorded during the game.

Background noises are in various types and at different volume levels during the game.

SAFE-T Speech Corpus
Approx. 300 hours of audio



1. Training Data Set
2. Development Data Set
3. Evaluation Data Sets
Reference keys, Metadata

The recorded audio and transcripts are organized to create the SAFE-T Speech Corpus.

The SAFE-T Speech Corpus package contains the three types of data sets for researchers.

Apply Speech Analytic To the Data

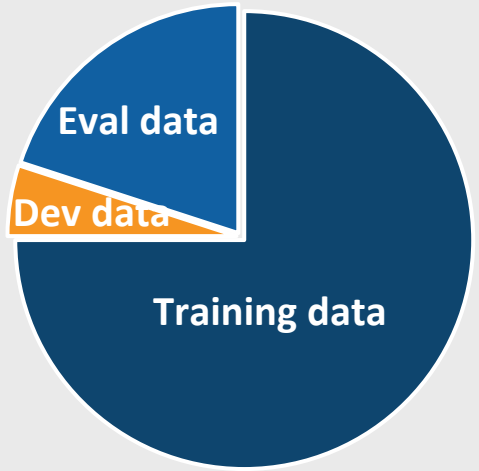
Audio Collected → **Data Sets Created** → **Audio Processed**

Communications are recorded during the game.

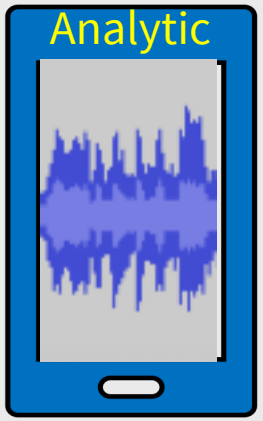
SAFE-T Speech Corpus
Approx. 300 hours of audio

Researchers join OpenSAT and obtain data sets

Researchers receive the data sets from LDC.



Audio →



Analytics (tasks):

→ speech activity detection
→ keyword search
→ speech-to-text

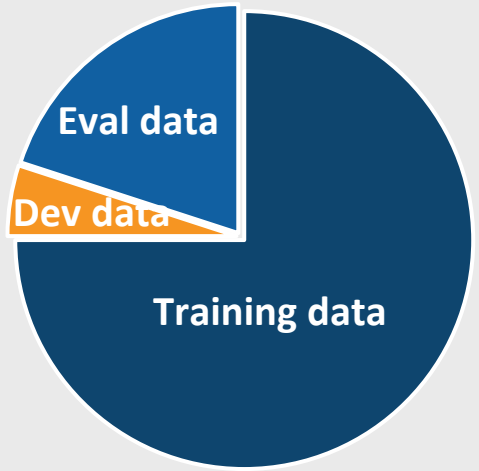
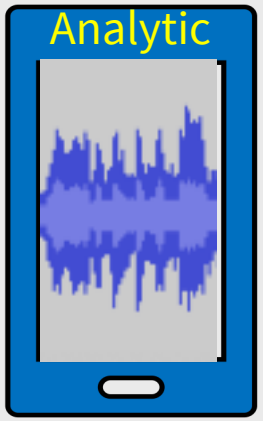
Background noises are in various types and at different volume levels during the game.

1. Training Data Set
2. Development Data Set
3. Evaluation Data Sets
Reference keys

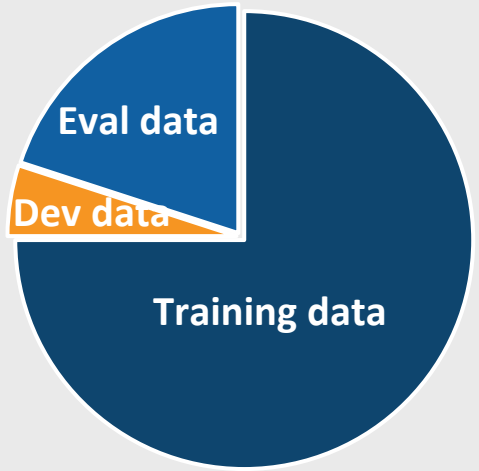
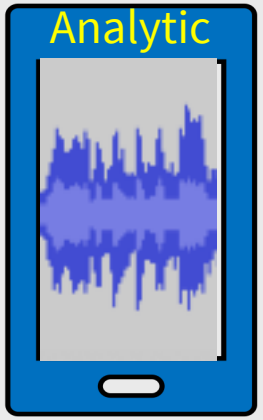
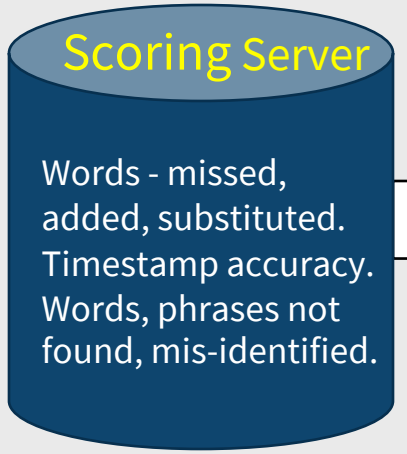
Task-specific analytic is applied to the data. The analytics are trained and developed for evaluation.

Analytic output is prepared for evaluation.

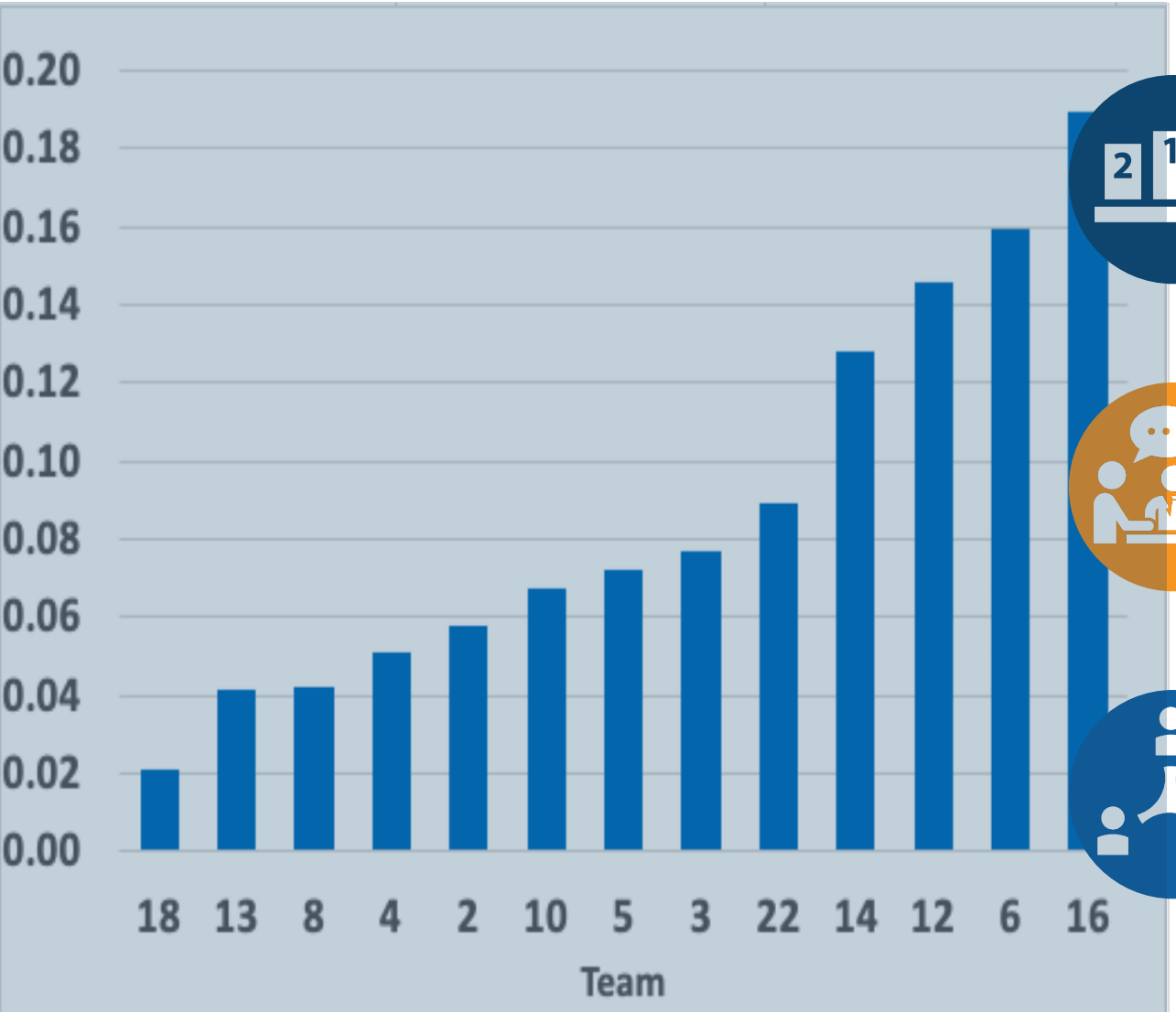
Analytic Output Is Evaluated

Audio Collected →	Data Sets Created →	Audio Processed →	Output Evaluated
Communications are recorded during the game.	SAFE-T Speech Corpus Approx. 300 hours of audio	Researchers apply their system to the audio	NIST scores system output with a scoring server
		Audio →  Output →	
Background noises are in various types and at different volume levels during the game.	1. Training Data Set 2. Development Data Set 3. Evaluation Data Sets Reference keys	Researchers upload their system's output to NIST	NIST ranks each system's score against the other systems' scores

Evaluation - Score

Audio Collected →	Data Sets Created →	Audio Processed →	Output Evaluated
Communications are recorded during the game.	SAFE-T Speech Corpus Approx. 300 hours of audio	Researchers apply their system to the audio	NIST scores system output with a scoring server
		Audio →  Output →	 → Score
Background noises are in various types and at different volume levels during the game.	1. Training Data Set 2. Development Data Set 3. Evaluation Data Sets Reference keys	Researchers upload their system's output to NIST	NIST ranks each system's score against the other systems' scores

Scores, Ranking, Workshop, Knowledge Transfer



Scores & Ranking

Researchers see how their system's performance compares to others.

Workshop

Researchers learn how the top ranked systems work from presentations and Q&A.

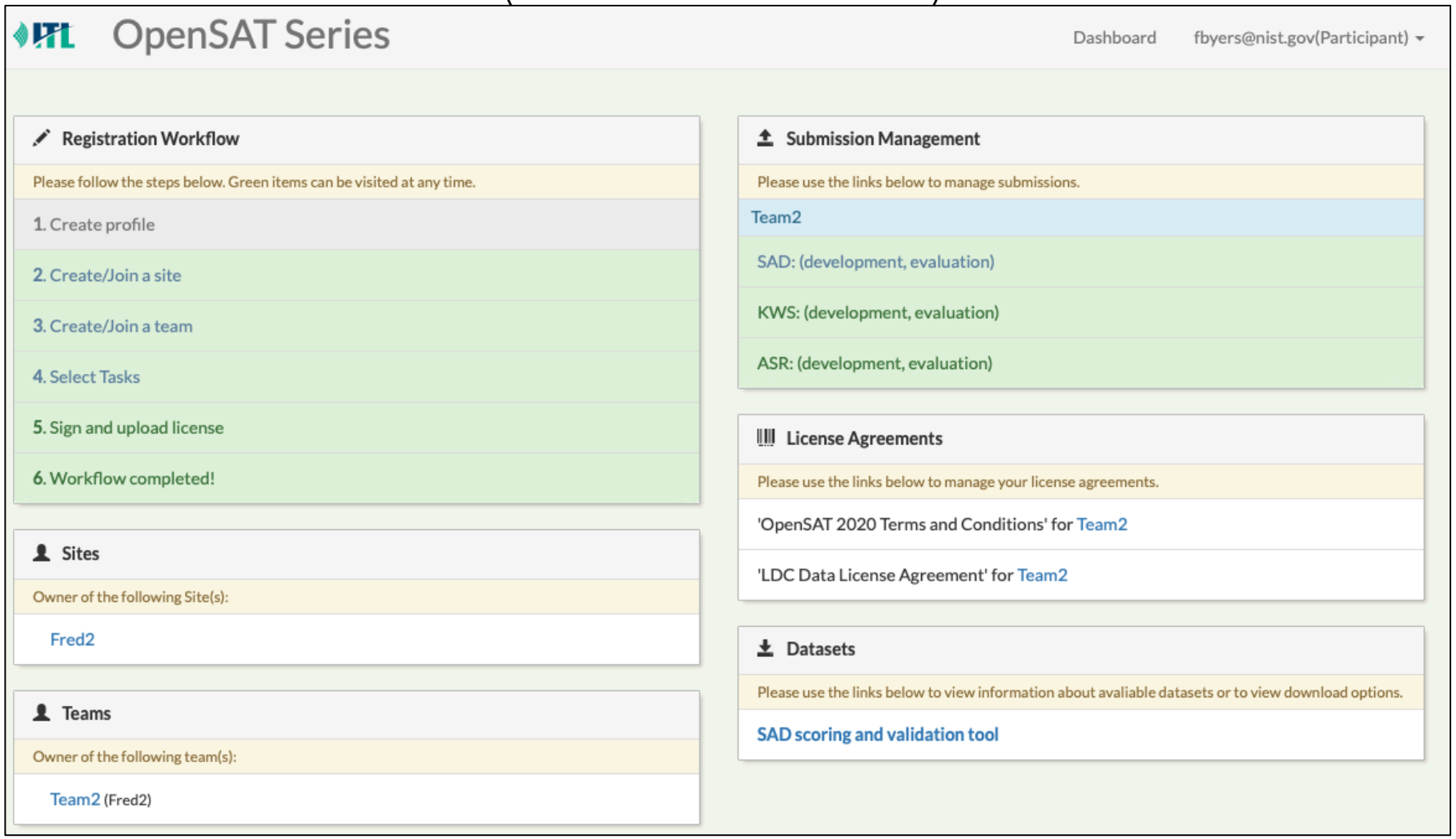
Technology Transfer

Researchers bring their knowledge to the industry and system development efforts.

OpenSAT20 Web Site

OpenSAT Web Site
Researcher Dashboard

(Screen Shot of Dashboard)



OpenSAT Evaluation Plan
Provides All The Details

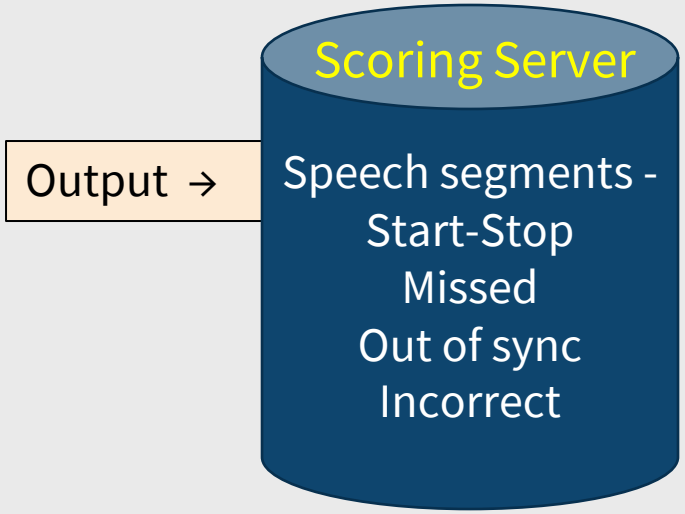
NIST Open Speech Analytic Technologies
2020 Evaluation Plan
(OpenSAT20)

Section 2

Equations and 2019 Test Results

Speech Activity Detection Scoring

Speech Activity Detection (SAD) Output Evaluated



$$DCF\ (\theta) = 0.75 \times P_{FN}\ (\theta) + 0.25 \times P_{FP}\ (\theta)$$

Detection Cost Function (DCF)

System goal is to detect the start and end of when a person speaks.

Scoring server calculates error probabilities from missed speech detection time and incorrect detection time.

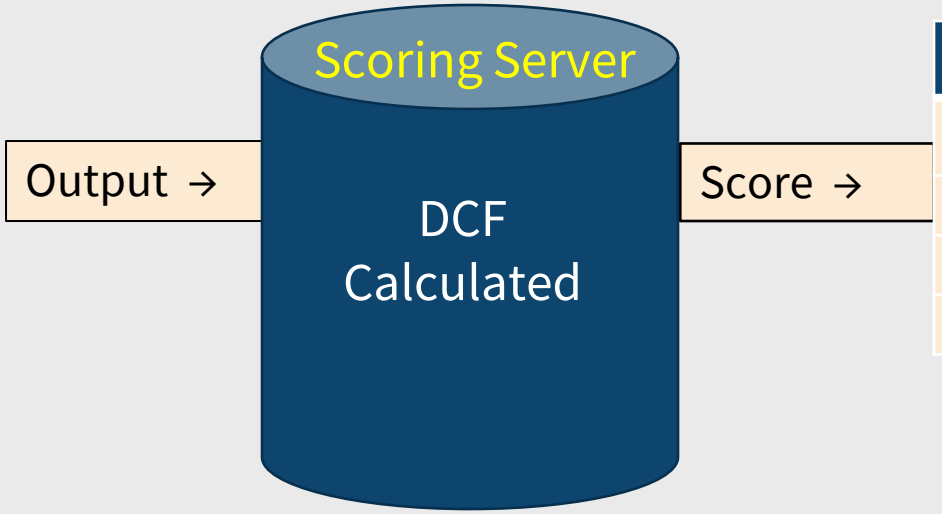
P_{FN} – Probability of a system missing speech
P_{FP} – Probability of a system incorrectly detecting speech

$$P_{FN} = \frac{\text{total missed speech time}}{\text{total actual speech time}}$$

$$P_{FP} = \frac{\text{total incorrect speech time}}{\text{total actual nonspeech time}}$$

Speech Activity Detection Scores

Speech Activity Detection (SAD) Output Evaluated



$$DCF\ (\theta) = 0.75 \times P_{FN}\ (\theta) + 0.25 \times P_{FP}\ (\theta)$$

Detection Cost Function (DCF)

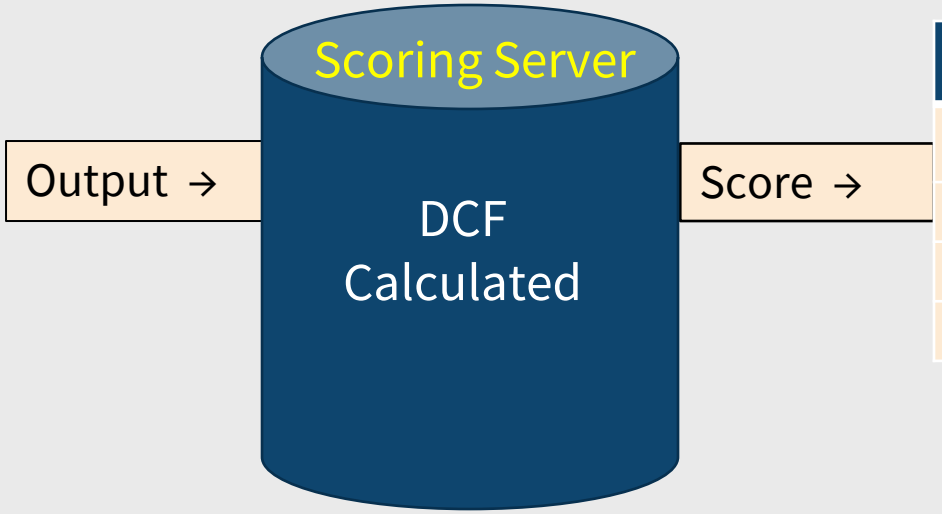
System scores for both loud and quiet background noise levels

System	DCF (Loud BG)	DCF (Quiet BG)	DCF Average
1	0.0481	0.0624	0.0525
2	0.1093	0.1175	0.1099
3	0.1479	0.0854	0.1245
4	0.1492	0.0911	0.1352

0.0 ---------- 1.0
Perfect DCF Bad

Speech Activity Detection Scores

Speech Activity Detection (SAD) Output Evaluated



$$DCF\ (\theta) = 0.75 \times P_{FN}\ (\theta) + 0.25 \times P_{FP}\ (\theta)$$

Detection Cost Function (DCF)

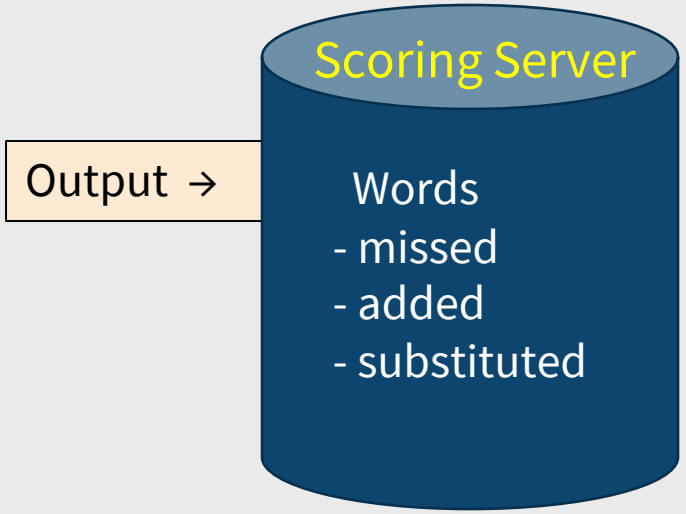
Individual speakers' scores with loud background noise level

Speaker	FN (Prob. Miss)	FP (Prob. Incorrect)	DCF
1	0.8992	0.9751	0.9181
2	0.0443	0.0345	0.0419
3	0.1763	0.0745	0.1508
4	0.0229	0.0531	0.0305

0.0 ---------- 1.0
Perfect DCF Bad

Automatic Speech Recognition Scoring

Automatic Speech Recognition (ASR) Output Evaluated



$$\text{WER} = \frac{(\text{N}_{\text{Del}} + \text{N}_{\text{Ins}} + \text{N}_{\text{Subst}})}{\text{N}_{\text{Ref}}}$$

Word Error Rate (WER)

System goal is to transcribe all audible speech to readable text.

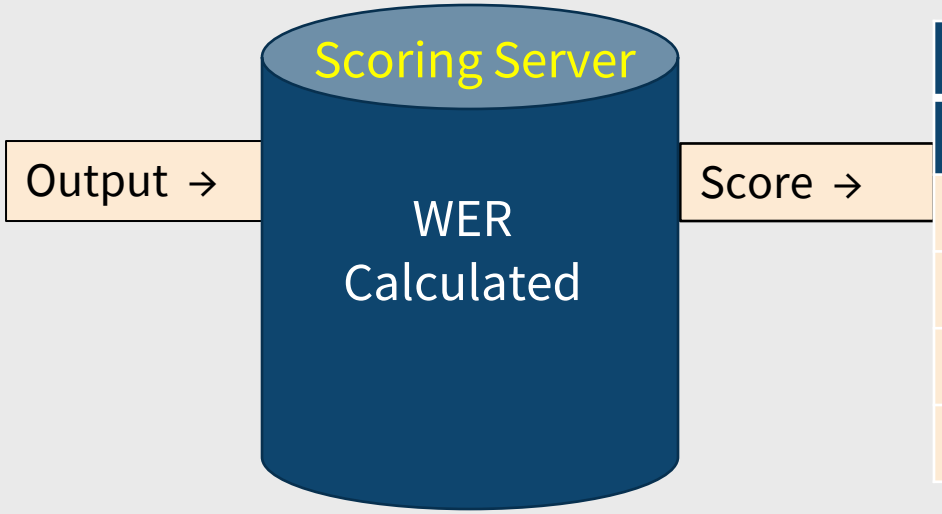
Scoring server detects and counts the number of missed words, inserted words, and mismatched words.

N_{Del} – Number of missed words
N_{Ins} – Number of inserted words (words added where there is no speaking)
N_{subst} – Number of non-matching words (e.g., misspelled, wrong word)

N_{Ref} – Total number of words in the transcript

Automated Speech Recognition Scores

Automatic Speech Recognition (ASR) Output Evaluated



$$\mathrm{WER} = \frac{(\mathrm{N_{Del}} + \mathrm{N_{Ins}} + \mathrm{N_{Subst}})}{\mathrm{N_{Ref}}}$$

Word Error Rate (WER)

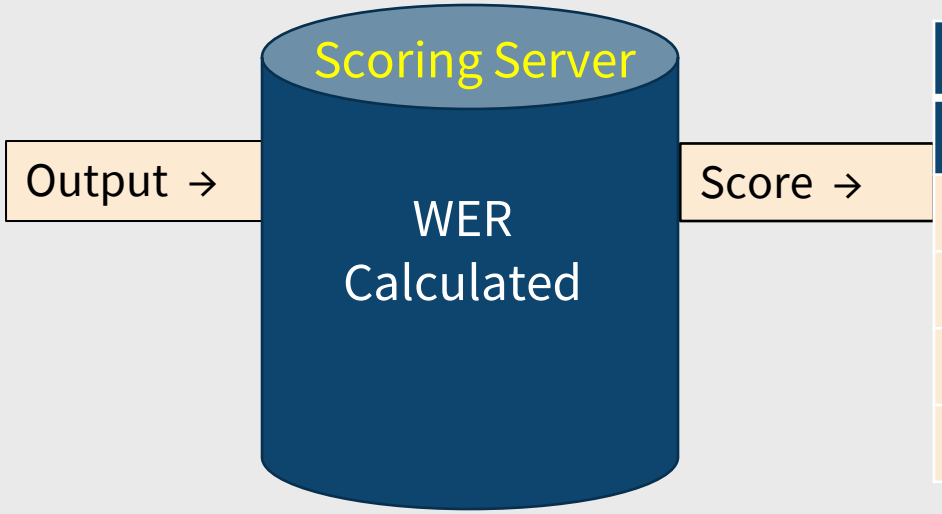
System scores

Loud and Quiet Backgrounds			
System	WER (Loud BG)	WER (Quiet BG)	Average WER
1	15.7%	15.0%	15.4%
2	36.1%	22.6%	29.3%

0.0% ---------- 100%
Perfect WER Bad

Automated Speech Recognition Scores

Automatic Speech Recognition (ASR) Output Evaluated



$$\text{WER} = \frac{(N_{Del} + N_{Ins} + N_{Subst})}{N_{Ref}}$$

Word Error Rate (WER)

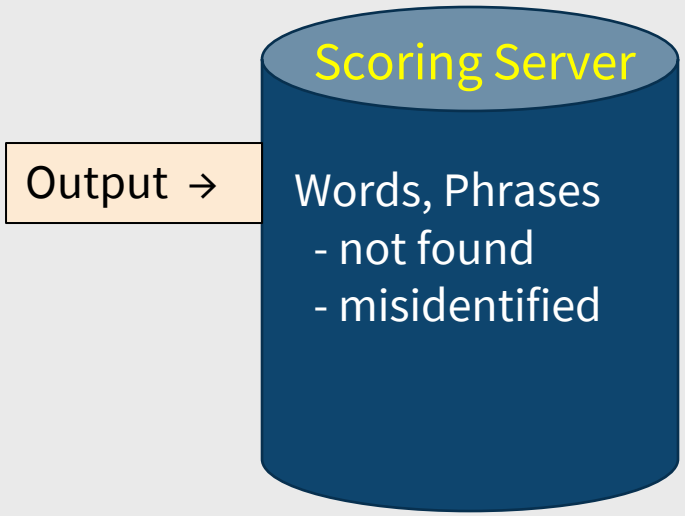
Speaker score examples for a system

WER with Loud Background					
Speaker	Total Words	Subst.	Missed	Inserted	WER
798	150	10	11	4	16.7%
798	308	18	112	5	43.8%
8801	131	7	16	1	18.3%
8801	227	29	46	0	33.0%

0.0% ---------- 100%
Perfect WER Bad

Keyword Search Scoring

Keyword Search (KWS) Output Evaluated



$$TWV(\theta) = 1 - [P_{FN}(\theta) + \beta \cdot P_{FP}(\theta)]$$

Term Weighted Value (TWV)

System goal is to automatically detect all keywords in the audio.

Scoring server calculates error probabilities from the number of missed keywords and incorrect detections.

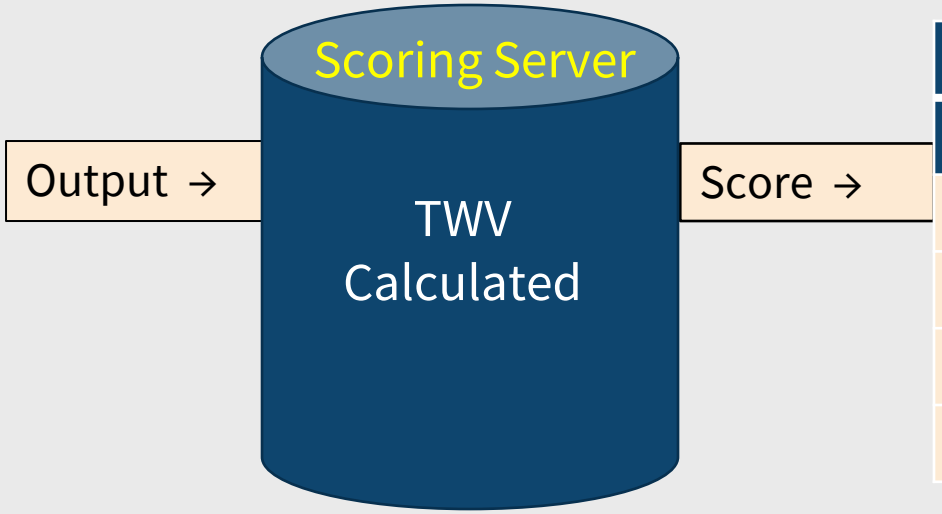
P_{FN} – Probability of a system missing a keyword
P_{FP} – Probability of a system detecting an incorrect word or an incorrect location

$$P_{FN} = \frac{\text{number of keywords missed (FN)}}{\text{total number of keywords}}$$

$$P_{FP} = \frac{\text{number of keywords misplaced (FP)}}{\text{total duration (seconds)}} \text{ / \#target words}$$

Keyword Search Scores

Keyword Search (KWS) Output Evaluated



$$TWV(\theta) = 1 - [P_{FN}(\theta) + \beta \cdot P_{FP}(\theta)]$$

Term Weighted Value (TWV)

Scores

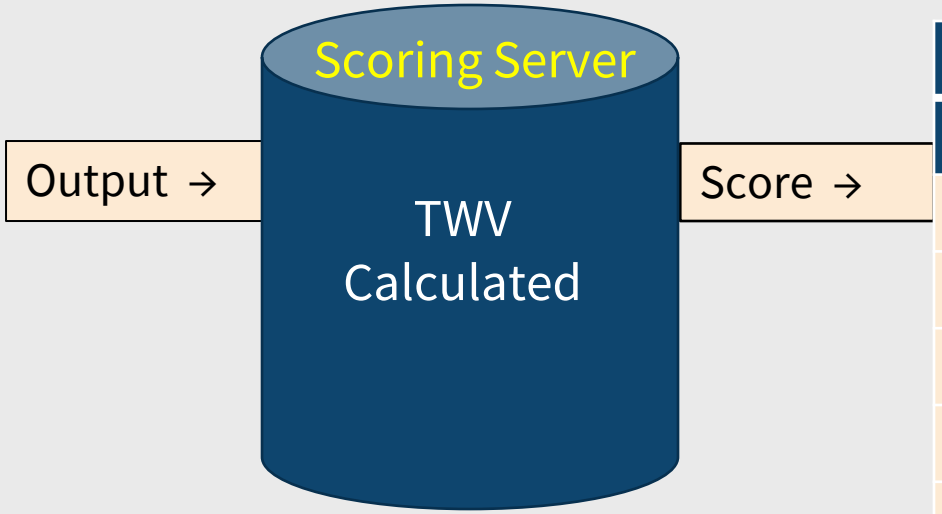
Loud and Quiet Backgrounds			
System	TWV (Loud BG)	TWV (Quiet BG)	TWV Average
1	0.2896	0.2455	0.2676
2	0.2525	0.3198	0.2862

------- 0.0 ---------- 1.0

Bad TWV Perfect

Keyword Search Scores

Keyword Search (KWS) Output Evaluated



$$TWV(\theta) = 1 - [P_{FN}(\theta) + \beta \cdot P_{FP}(\theta)]$$

FN=Miss

Term Weighted Value (TWV)

Scores

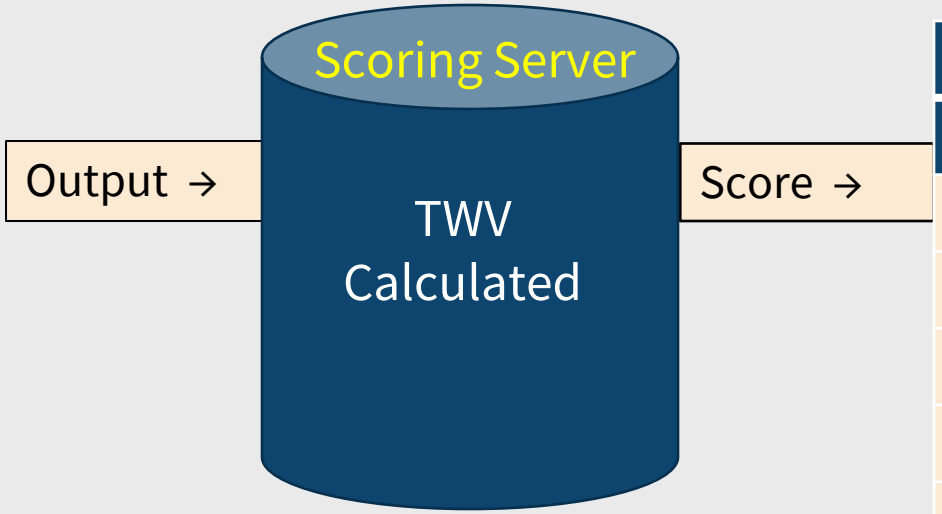
Examples for individual keywords						
Keyword	**Instances**	**Miss**	**FP**	**P_{Miss}**	**P_{FP}**	**TWV**
alarm	3	2	1	0.667	0.00011	0.2222
exit	7	4	1	0.571	0.00011	0.3174
building	6	1	0	0.167	0.00000	0.8333
standing	3	0	0	0.000	0.00000	1.0000
over	20	10	38	0.500	0.00848	-7.9813

------- 0.0 ---------- 1.0

Bad TWV Perfect

Keyword Search Scores

Keyword Search (KWS) Output Evaluated



$$TWV(\theta) = 1 - [P_{FN}(\theta) + \beta \cdot P_{FP}(\theta)]$$

FN=Miss

Term Weighted Value (TWV)

Scores

Keyword TWV example with Loud and Quiet Background						
Keyword	**Instances**	**Miss**	**FP**	**P_{Miss}**	**P_{FP}**	**TWV**
time (L)	7	3	1	0.429	0.00011	0.4602
time (Q)	5	2	1	0.400	0.00022	0.3776
time (L)	7	5	0	0.714	0.00000	0.2857
time (Q)	5	3	0	0.600	0.00000	0.4000

------- 0.0 ---------- 1.0

Bad TWV Perfect

Moving Forward

- **2020**
  - Complete OpenSAT20 Evaluation
  - Compare OpenSAT20 and OpenSAT19 results (4th quarter)
  - Obtain high-quality transcription of real-world operational audio

- **2021**
  - OpenSAT21
  - Continue with the LDC data
  - Include real-world operational data for testing
  - Add the speaker diarization task (keeping track of speakers A, B, C, etc.)

Contact Us

NIST
100 Bureau Dr, MS 894
Gaithersburg, MD 20899

OpenSAT Web Site https://sat.nist.gov
Group Web Site https://www.nist.gov/itl/iad/mig/opensat

NIST/ITL/ Information Access Division/Multimodal Information Group
Fred Byers frederick.byers@nist.gov Office: 301-975-2909


THANK YOU
#PSCR2020