Speech Analytics Evaluation Project (OpenSAT)

Fred Byers
NIST/Information Technology Laboratory
Gaithersburg, MD

#PSCR2020
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

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

<table>
<thead>
<tr>
<th>Task Description</th>
<th>Task Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. <strong>Speech Activity Detection</strong></td>
<td>Automatically detect the beginning and ending of speaking segments by timestamps.</td>
</tr>
<tr>
<td>2. <strong>Automatic Speech Recognition</strong></td>
<td>Automatically transcribe verbatim all audible speech to readable text.</td>
</tr>
<tr>
<td>3. <strong>Keyword Search</strong></td>
<td>Automatically detect all spoken instances of a predetermined list of words and phrases.</td>
</tr>
</tbody>
</table>

- **Task Example 1:** Going inside now
- **Task Example 2:** Going inside now
- **Task Example 3:** Going inside now

- **Task Example 1:** Going inside now
- **Task Example 2:** Going inside now
- **Task Example 3:** Going inside now
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) recorded and transcribed dozens of fire rescue gameplaying 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.
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.
Communications are recorded during the game. Background noises are in various types and at different volume levels during the game.

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.

1. Training Data Set
2. Development Data Set
3. Evaluation Data Sets
Reference keys, Metadata
Apply Speech Analytic To the Data

Audio Collected
- Communications are recorded during the game.
- Background noises are in various types and at different volume levels during the game.

Data Sets Created
- SAFE-T Speech Corpus
- Approx. 300 hours of audio
- 1. Training Data Set
- 2. Development Data Set
- 3. Evaluation Data Sets
- Reference keys

Audio Processed
- Researchers join OpenSAT and obtain data sets
- Researchers receive the data sets from LDC.

Analytics (tasks):
- speech activity detection
- keyword search
- speech-to-text

Analytic output is prepared for evaluation.
Communications are recorded during the game.

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 apply their system to the audio

Researchers upload their system’s output to NIST

NIST ranks each system’s score against the other systems’ scores
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

Researchers apply their system to the audio

Researchers upload their system’s output to NIST
- NIST scores system output with a scoring server
- 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.
Welcome to the NIST Open Speech Analytic Technologies Evaluation (OpenSAT20)

OpenSAT20 will be supporting the public safety communications domain.

OpenSAT will be hosting the Linguistic Data Consortium (LDC) DIHARD III in 2020. Go to the DIHARD III tab for more information.

Summary

OpenSAT20 is the second in the OpenSAT Series for speech analytic systems evaluations. OpenSAT provides an opportunity for participants to compare their system performance against a pool of systems performances for each task and is intended to encourage cross-learning among developers.

Tasks and Domains

<table>
<thead>
<tr>
<th>System Tasks</th>
<th>Data Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic Speech Recognition (ASR)</td>
<td>Public safety communications (PSC)</td>
</tr>
<tr>
<td>Speech Activity Detection (SAD)</td>
<td></td>
</tr>
<tr>
<td>Keyword Search (KWS)</td>
<td></td>
</tr>
</tbody>
</table>
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)

1 Introduction................................................................................................................. 2
2 Planned Schedule ........................................................................................................ 2
3 Data............................................................................................................................. 3
4 Tasks - Overview and Performance Metrics .................................................................. 4
5 Evaluation Rules ......................................................................................................... 4
6 Evaluation Protocol ..................................................................................................... 5
7 System Input, Output, and Performance Metrics ............................................................ 5
Appendix I  SAD ............................................................................................................ 7
Appendix II  KWS ......................................................................................................... 10
Appendix III ASR .......................................................................................................... 15
Appendix IV SAD, KWS, and ASR - System Output Submission Packaging .................... 18
Appendix V SAD, KWS, and ASR - System Descriptions and Auxiliary Condition Reporting . 19
Section 2

Equations and 2019 Test Results
Speech Activity Detection Scoring

**Speech Activity Detection (SAD)**
Output Evaluated

**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

**Detection Cost Function (DCF)**

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

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

\[ P_{FP} = \frac{\text{total incorrect speech time}}{\text{total actual nonspeech time}} \]
Speech Activity Detection Scores

**Speech Activity Detection (SAD) Output Evaluated**

Output → DCF Calculated → Score

**Detection Cost Function (DCF)**

System scores for both loud and quiet background noise levels

<table>
<thead>
<tr>
<th>System</th>
<th>DCF (Loud BG)</th>
<th>DCF (Quiet BG)</th>
<th>DCF Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0481</td>
<td>0.0624</td>
<td>0.0525</td>
</tr>
<tr>
<td>2</td>
<td>0.1093</td>
<td>0.1175</td>
<td>0.1099</td>
</tr>
<tr>
<td>3</td>
<td>0.1479</td>
<td>0.0854</td>
<td>0.1245</td>
</tr>
<tr>
<td>4</td>
<td>0.1492</td>
<td>0.0911</td>
<td>0.1352</td>
</tr>
</tbody>
</table>

DCF (θ) = 0.75 × P_{FN} (θ) + 0.25 × P_{FP} (θ)

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

DFC (θ) = 0.75 × P_{FN} (θ) + 0.25 × P_{FP} (θ)
Speech Activity Detection Scores

Speech Activity Detection (SAD) Output Evaluated

Detection Cost Function (DCF)

Individual speakers’ scores with loud background noise level

<table>
<thead>
<tr>
<th>Speaker</th>
<th>FN (Prob. Miss)</th>
<th>FP (Prob. Incorrect)</th>
<th>DCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8992</td>
<td>0.9751</td>
<td>0.9181</td>
</tr>
<tr>
<td>2</td>
<td>0.0443</td>
<td>0.0345</td>
<td>0.0419</td>
</tr>
<tr>
<td>3</td>
<td>0.1763</td>
<td>0.0745</td>
<td>0.1508</td>
</tr>
<tr>
<td>4</td>
<td>0.0229</td>
<td>0.0531</td>
<td>0.0305</td>
</tr>
</tbody>
</table>

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

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

Perfect | DCF | Bad
Automatic Speech Recognition (ASR) Output Evaluated

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

Word Error Rate (WER)

$$WER = \frac{(N_{Del} + N_{Ins} + N_{subst})}{N_{Ref}}$$
Automated Speech Recognition Scores

Word Error Rate (WER)

System scores

<table>
<thead>
<tr>
<th>System</th>
<th>WER (Loud BG)</th>
<th>WER (Quiet BG)</th>
<th>Average WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.7%</td>
<td>15.0%</td>
<td>15.4%</td>
</tr>
<tr>
<td>2</td>
<td>36.1%</td>
<td>22.6%</td>
<td>29.3%</td>
</tr>
</tbody>
</table>

Automatic Speech Recognition (ASR)
Output Evaluated

Scoring Server

WER Calculated

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

Output →

Score →

0.0% ------------------------------- 100%
Perfect                          WER                          Bad
Automated Speech Recognition Scores

Automatic Speech Recognition (ASR) Output Evaluated

Word Error Rate (WER)

Speaker score examples for a system

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Total Words</th>
<th>Subst.</th>
<th>Missed</th>
<th>Inserted</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>798</td>
<td>150</td>
<td>10</td>
<td>11</td>
<td>4</td>
<td>16.7%</td>
</tr>
<tr>
<td>798</td>
<td>308</td>
<td>18</td>
<td>112</td>
<td>5</td>
<td>43.8%</td>
</tr>
<tr>
<td>8801</td>
<td>131</td>
<td>7</td>
<td>16</td>
<td>1</td>
<td>18.3%</td>
</tr>
<tr>
<td>8801</td>
<td>227</td>
<td>29</td>
<td>46</td>
<td>0</td>
<td>33.0%</td>
</tr>
</tbody>
</table>

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

0.0% .......................... 100%  
Perfect  ....................... Bad
Keyword Search **Scoring**

**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

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

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

**Output →**

**Scoring Server**

**Words, Phrases**
- not found
- misidentified

**Keyword Search (KWS)**

**Output Evaluated**

**Term Weighted Value (TWV)**

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

#### Term Weighted Value (TWV)

**Scores**

<table>
<thead>
<tr>
<th>System</th>
<th>TWV (Loud BG)</th>
<th>TWV (Quiet BG)</th>
<th>TWV Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2896</td>
<td>0.2455</td>
<td>0.2676</td>
</tr>
<tr>
<td>2</td>
<td>0.2525</td>
<td>0.3198</td>
<td>0.2862</td>
</tr>
</tbody>
</table>

**TWV (θ)**:

\[ TWV (θ) = 1 - [P_{FN} (θ) + β \cdot P_{FP} (θ)] \]

---

**Evaluation Range**

0.0 to 1.0

- **Bad**
- **TWV**
- **Perfect**
Keyword Search Scores

Keyword Search (KWS) Output Evaluated

Term Weighted Value (TWV) Scores

Examples for individual keywords

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Instances</th>
<th>Miss</th>
<th>FP</th>
<th>( P_{\text{Miss}} )</th>
<th>( P_{\text{FP}} )</th>
<th>TWV</th>
</tr>
</thead>
<tbody>
<tr>
<td>alarm</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0.667</td>
<td>0.00011</td>
<td>0.2222</td>
</tr>
<tr>
<td>exit</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>0.571</td>
<td>0.00011</td>
<td>0.3174</td>
</tr>
<tr>
<td>building</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0.167</td>
<td>0.00000</td>
<td>0.8333</td>
</tr>
<tr>
<td>standing</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>0.00000</td>
<td>1.0000</td>
</tr>
<tr>
<td>over</td>
<td>20</td>
<td>10</td>
<td>38</td>
<td>0.500</td>
<td>0.00848</td>
<td>-7.9813</td>
</tr>
</tbody>
</table>

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

FN=Miss

\[ \text{TWV} = \begin{cases} 1 & \text{Perfect} \\ 0.0 & \text{Bad} \end{cases} \]
Keyword Search Scores

Keyword Search (KWS)
Output Evaluated

Term Weighted Value (TWV)

Scores

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Instances</th>
<th>Miss</th>
<th>FP</th>
<th>$P_{Miss}$</th>
<th>$P_{FP}$</th>
<th>TWV</th>
</tr>
</thead>
<tbody>
<tr>
<td>time (L)</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>0.429</td>
<td>0.00011</td>
<td>0.4602</td>
</tr>
<tr>
<td>time (Q)</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>0.400</td>
<td>0.00022</td>
<td>0.3776</td>
</tr>
<tr>
<td>time (L)</td>
<td>7</td>
<td>5</td>
<td>0</td>
<td>0.714</td>
<td>0.00000</td>
<td>0.2857</td>
</tr>
<tr>
<td>time (Q)</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0.600</td>
<td>0.00000</td>
<td>0.4000</td>
</tr>
</tbody>
</table>

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

FN=Miss

---

0.0 -------------------------------------------------- 1.0
Bad TWV Perfect
Moving Forward

• 2020
  • Complete OpenSAT20 Evaluation
  • Compare OpenSAT20 and OpenSAT19 results (4\textsuperscript{th} 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

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

NIST
100 Bureau Dr, MS 894
Gaithersburg, MD 20899
THANK YOU

#PSCR2020