NISTIR 8242

2017 Pilot Open Speech Analytic Technologies Evaluation (2017 NIST Pilot OpenSAT)

Post Evaluation Summary

Fred Byers
Omid Sadjadi

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8242

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8242

2017 Pilot Open Speech Analytic Technologies Evaluation (2017 NIST Pilot OpenSAT)

Post Evaluation Summary

Fred Byers
Information Access Division
Information Technology Laboratory

Omid Sadjadi
Systems Plus Inc.

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8242

March 2019

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

Abstract

NIST conducted the 2017 Pilot Open Speech Analytic Technologies Evaluation (OpenSAT) Evaluation as the first in a new type of evaluation series. This new series is being designed to combine data domains and speech-analytic tasks including the public-safety-communications domain. This report summarizes the set up and results of the Pilot. The report includes 1) anonymous, composite plots, 2) an analysis of systems output, and 3) the dataset challenges in the evaluation. The methods used in the Pilot OpenSAT Evaluation included speech-activity detection; keyword search; automatic, speech recognition; and, three data domains. Those domains were low-resource language, public-safety communications, and YouTube videos. For the speech-analytics task, participants in the evaluation could choose one or all of the methods and one or all of the data domains.

Key words

Automatic speech recognition; keyword search; speech activity detection; voice activity detection; evaluation; public safety communications; Babel; speech analytic system; speech analytic task; Lombard effect.

Institutional Review Board

The National Institute of Standards and Technology (NIST) Human Subjects Protection Office (HSPO) reviewed the protocol for this project (ITL-17-0011) and determined it is not human subjects research as defined in Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule for the Protection of Human Subjects (45 CFR 46, Subpart A).

Disclaimer

Any mention of commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by NIST nor does it imply that the products mentioned are necessarily the best available for the purpose.

Table of Contents

1 Introduction *1*

1.1 Background 1

Schedule 2

2 Tasks *2*

2.1 SAD Task 2

2.2 KWS Task 3

2.3 ASR Task 5

3 Data *5*

3.1 PSC Data 5

3.2 VAST Data 6

3.3 Babel Data 6

4 Participants *7*

5 Results *8*

5.1 SAD Results 8

5.2 KWS Results 10

5.3 ASR Results 11

6 Analysis and Review *12*

6.1 SSSF Dataset Review 12

6.2 VAST Dataset Review 12

6.2.1 Files from the VAST development set reviewed – First method. 13

6.2.2 Files from the VAST evaluation set reviewed – Second method 15

6.3 Babel Dataset Review 20

7 Summary *20*

8 References *22*

List of Tables

Table 1. Data domains and tasks for the Pilot OpenSAT Evaluation 2
Table 2. Table shows tasks performed and the data domains used by each team 7
Table 3. Ten worst scores of the seven best scoring teams 16
Table 4. VVC010551 audio file review 16
Table 5. VVC024151 audio file review 17
Table 6. VVC009090 audio file review 18
Table 7. VVC013898 audio file review 19
Table 8. VVC037067 audio file review 19

List of Figures

Figure 1. Comparing a hypothetical system-detection against a reference annotation............. 3
Figure 2. SSSF Data, SAD Results 8
Figure 3. VAST Data, SAD Results 9
Figure 4. Babel Data, SAD Results 9
Figure 5. Contrast of DCF scores for SAD across the three data domains............................. 10
Figure 6. Babel Data, KWS Results.. 10
Figure 7. SSSF Data, KWS Results... 11
Figure 8. Babel Data, ASR Results 11
Figure 9. SSSF Data, ASR Results 11
Figure 10. VVC011246 audio file 13
Figure 11. VVC031565 audio file 13
Figure 12. VVC004451 audio file 14
Figure 13. VVC026043 audio file 14
Figure 14. VVC009379 audio file 15
Figure 15. VVC041917 audio file 15

1 Introduction

1.1 Background

The Pilot Open Speech Analytic Technologies (OpenSAT) Evaluation was intended as an initial proof-of-concept for a new Evaluation Series. This report focuses on the pilot only. It includes an overview of the framework, multi-task approach, data, data analysis and review, prediction measures that were implemented in the OpenSAT Evaluation, and a follow up summary discussion.

A primary goal of this new series is to provide a comprehensive method for assessing the performance of targeted speech analytic technologies for public safety communications. Organizations responsible for public safety need conclusive, science-based, assessment methods to assess the performance of these technologies **before** they become part of these safety critical systems and to ensure the best performance possible for voice communications in their unique environment and for processing of communications after an event. Before this pilot, a well-defined performance assessment methodology for speech analytic systems where the Lombard effect[1] and stressed communications within a variety of background scenarios, such as in first responder public safety environments, did not exist. To address these challenges, after the Pilot, successive evaluations in the series are intended to include controlled, **simulated**, public-safety related communications.

Another goal of both the pilot and the series is to accelerated the development of speech-analytic systems by increasing opportunities for information sharing among technology developers. This will be achieved by 1) combining tasks in an evaluation that are typically evaluated independently and 2) designing the series to leverage such multi-task evaluations by including multiple data domains.

The OpenSAT framework is designed to be an online evaluation using a NIST webserver for managing registration, data-license agreements, reference files, tools, system-output uploads, and summary plots. Access to data in the Pilot OpenSAT was made available through cooperation with the Linguistic Data Consortium (LDC) at UPENN for managing license agreements, cataloging and the releasing of audio files and metadata to developers.

Three speech analytic tasks and three speech domains were included. The three tasks were **Speech Activity Detection (SAD), Automatic Speech Recognition (ASR), and Keyword Search (KWS)**. The three speech domains included a low resource language domain (Pashto language telephone conversations from the Intelligence Advanced Research Projects Activity (IARPA) **Babel data set**) [1], YouTube video domain (audio extracted from a variety of YouTube amateur videos to create the Video Annotation for Speech Technologies **(VAST**) **data set**), and the Public Safety Communications domain (mobile radio and telephone dispatches to create a Sofa Super Store Fire **(SSSF) data set**).

[1] https://en.wikipedia.org/wiki/Lombard_effect

Table 1. Data domains and tasks for the Pilot OpenSAT Evaluation

Domain	Task	Language
Public Safety Communications (Sofa Super Store Fire dispatches)	SAD, ASR, KWS	English
VAST Collection (YouTube Videos)	SAD	Arabic, Mandarin, English
IARPA Babel - Low Resource Language (Pashto language)	SAD, ASR, KWS	Pashto

1.2 Schedule

April 26 through May 30, 2017 - Registration period
April 26, 2017 - Development data released to participants
May 30, 2017 - Evaluation data released and system output submissions to NIST opens
June 20, 2017 - System output submission to NIST closes
July 14, 2017 - System output results released
July 14, 2017 - Evaluation reference data released
July 21, 2017 - NIST scoring server active again for optional continued system development
Sept. 6, 2017 - Teleconference with performers to discuss Pilot performance/experience

2 Tasks

2.1 SAD Task

Speech-analytic systems are expected to detect the presence of **all** speech occurrences, or speech segments, in audio recordings automatically. Audio recordings are of variable durations and with variable speech-to-non-speech ratios. A SAD system is scored by comparing its system-identified start & end times output for all speech in audio recordings to human-identified start & end times for those recordings. Correct, incorrect, and partially correct results determine error probabilities for the system's performance for each domain dataset [2]. Four possibilities were considered when evaluating SAD systems output for all data domains.

1. True Positive (**TP**) - system correctly identifies start-stop times of speech segments compared to the reference (manual annotation),
2. True Negative (**TN**) - system correctly identifies start-stop times of non-speech segments compared to reference,
3. False Positive (**FP**), (False Alarm) - system incorrectly identifies speech in a segment where the reference identifies the segment as non-speech,
4. False Negative (**FN**), (False Reject) - system missed identification of speech in a segment where the reference identifies a segment as speech.

The evaluation metric for SAD system performance is the detection cost function (DCF). The value of this function (see Eq 1) is a probabilistic measure of a system's false negative responses (missed detections) and the false positive responses (false alarm). These responses are based on the system's output for each audio file. The values for all files within the target data set (development or evaluation) are then combined and averaged to provide an overall system score.

$$DCF\ (\theta) = 0.75 \times P_{\text{Miss}}\ (\theta) + 0.25 \times P_{\text{FA}}\ (\theta) \qquad \text{(Eq 1)}$$

θ – denotes a given system decision threshold setting. System developers determine a system's speech detection threshold (θ) setting for their systems with the goal of minimizing the DCF value. Missed detections are considered more critical than false alarms and as a result the probability for misses is penalized/weighted 3 times more than the probability for false alarms. The weighting is consistent with historical weighting in DCF calculation for previous evaluations conducted by NIST[2].

Figure 1 shows the four possibilities when comparing system detected results against the reference annotation. The 0.5s collar is a 0.5-second non-scored area "buffer zone" immediately preceding and following each speech segment in the reference annotation The FN probability is determined by dividing the system output total FN time by the reference total speech time for an audio file, and the FP probability is determined by dividing the system output total FP time by the reference total non-speech time. These probabilities are used to calculate the DCF value for each audio file.

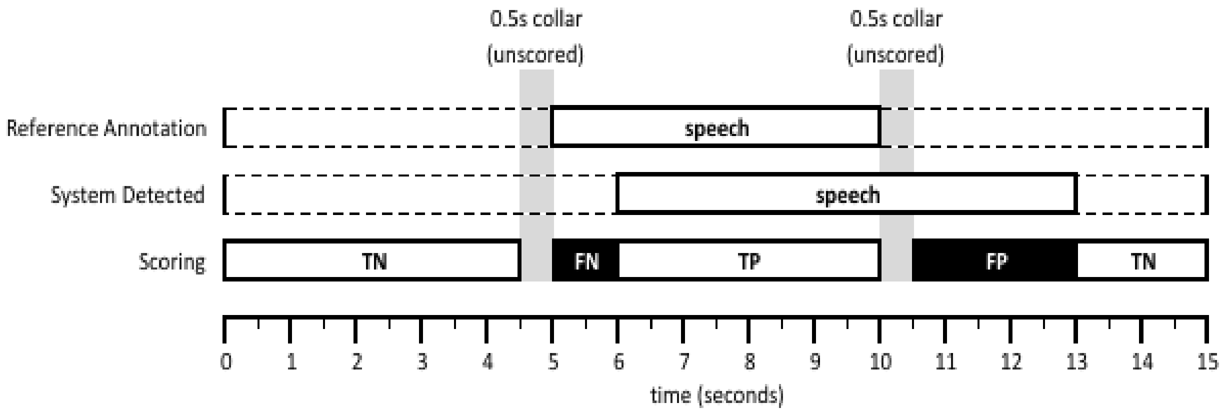


Figure 1. Comparing a hypothetical system-detection against a reference annotation

2.2 KWS Task

The goal of the KWS task is to determine if a speech-analytic system can detect all occurrences of a "keyword" automatically. A keyword in an audio recording is a pre-defined single word or phrase, which is transcribed with the spelling convention used in a language's original orthography. Each such system-detected keyword shall have the correct spelling and have both the beginning and end time-stamps [3].

[2] e.g., NIST Open Evaluation of Speech Activity Detection (OpenSAD15) https://www.nist.gov/sites/default/files/documents/itl/iad/mig/Open_SAD_Eval_Plan_v10.pdf

Keyword detection performance will be measured as a function of Missed Detection and False Alarm error types. Four system output possibilities are considered for scoring key-word regions.

1. (TP) - correct system detection of a keyword (matches the reference location and spelling)
2. (TN) - correct system non-detection of a keyword where a keyword does not exist
3. (FN) or (Miss) – system non-detection or misspelling of a keyword
4. (FP) or (FA) – system detection of a keyword not in the reference or correct location

Scoring protocol will be the "Keyword Occurrence Scoring" protocol that evaluates system accuracy based on the three steps below.

1. Reference-to-system keyword alignment
   - The KWS evaluation uses the Hungarian Solution to the Bipartite Graph matching problem[3] to compute the minimal cost for 1:1 alignment (mapping) of reference keywords to system output keywords.

2. Performance metric computation Term Weighted Value (TWV), Actual Term Weighted Value (ATWV)
   - Uses probability values derived for FP (or FA), and Miss (or FN).
   - System Actual TWV (ATWV): a measure of keyword detection performance at a given system's threshold setting (θ).
   - System Maximum TWV (MTWV): an oracle measure of keyword detection performance at the system's optimal θ setting. (The difference between ATWV and MTWV indicates the loss in performance due to a less-than-optimal system threshold (θ) setting for ATWV when determining the θ for ATWV.)

3. Detection Error Tradeoff (DET) Curves[4]
   - Curve depicts the tradeoff between missed detections versus false alarms for a range of θ settings.

$$TWV(\theta) = 1 - [P_{Miss}(\theta) + \beta \cdot P_{FA}(\theta)] \qquad \text{(Eq 2)}$$

Choosing θ:

- Developers choose a decision threshold for their "Actual Decisions" to optimize their term-weighted value: All the "YES" system occurrences
- The score obtained using this threshold is called the "**Actual Term Weighted Value**" (ATWV)
- The evaluation code searches for the system's optimum decision score threshold
- The score obtained using this method is called the "**Maximum Term Weighted Value**" (MTWV)

[3] Harold W. Kuhn, "The Hungarian Method for the assignment problem", *Naval Research Logistic Quarterly*, **2**:83-97, 1955

[4] https://en.wikipedia.org/wiki/Detection_error_tradeoff

2.3 ASR Task

The goal of the ASR system is to automatically detect all words in an audio recording and produce verbatim, a case-insensitive transcript of all the words spoken in that audio. ASR task outputs a stream of Conversation Time Marked (CTM) lexical tokens. That stream reports 1) the token's begin & duration times within the recording, 2) the spelling of the token, and 3) a confidence score, which is a value in the range [0,1], indicating the system's confidence that the token is correct [3]. Four system output possibilities were considered.

1. Correct output - system correctly locates and correctly spells a lexical token item (token) compared to the reference lexical token location and spelling,
2. Missed/Deleted output - system output misses the detection of a reference lexical token,
3. Inserted output - system outputs a lexical token where it does not exist (no mapping) in the reference,
4. Substitution - system output correctly locates but miss-spells a lexical token compared to the spelling of the reference token.

System scoring involved three steps: 1) normalization of system output tokens, 2) system output tokens aligned with reference tokens, and 3) system output performance computation. System output performance is a computation of the Word Error Rate (WER). The computation is the fraction of token recognition errors per maximum number of reference tokens. Below is the formula.

$$\text{WER} = \frac{(N_{Del} + N_{Ins} + N_{Subst})}{N_{Ref}} \qquad \text{(Eq 3)}$$

where
N_{Del} = number of unmapped reference tokens (tokens not detected by the system)
N_{Ref} = the maximum number of reference tokens (includes scorable and optionally deletable reference tokens)
N_{Ins} = number of unmapped system outputs tokens (tokens that are not in the reference)
N_{Subst} = number of system output tokens mapped to reference tokens but non-matching to the reference spelling

3 Data

3.1 PSC Data

Sofa Super Store Fire (SSSF) Dispatches [4]

A data set was created from the audio of, and logs of, radio and telephone dispatches from the Sofa Super Store Fire that occurred June 18, 2007 in Charleston, South Carolina [4]. These dispatches had previously been made available from the Charleston Sofa Super Store Phase II report (published May 15, 2008) by the City of Charleston and through the FOIA process. The report contains textual transcription of those dispatches. The transcriptions were re-annotated and transformed by NIST into the formats required to provide a reference key for scoring system's output in the Pilot OpenSAT evaluation. The resulting data set consists of approximately one hour total of audio and transcription from the dispatches for SAD, KWS, and ASR. The approximately one hour audio is divided into twelve roughly five-minute audio files, six for system development and six for system evaluation, and their respective transcriptions.

The recorded audio represents real-world, fire-response, operational data that cannot be duplicated through a controlled scientific experiment or simulation. The data presents multiple challenges for system's analytics such as land-mobile-radio transmission effects, speaking with significant background noise (Lombard effect), speech under cognitive and physical stress, varying background noise types, varying background decibel levels, and a real-world scenario.

3.2 VAST Data

Audio extracted from YouTube Videos

A subset of the VAST audio dataset was created specifically and only for the SAD analytic task. It was created by extracting audio from a subset of a YouTube video collection which was created by the Linguistic Data Consortium (LDC) for NSA research. A total of 500 hours of audio were extracted from the video collection: 300 for system development and 200 for system evaluation. The development set consisted of a total of 13.31 hours of audio, with approximately 11.04 hours for speech and 2.26 hours for non-speech. The length of each file ranged from approximately 20 seconds to 5 minutes. Languages spoken in audio files included Arabic, Mandarin and English. Annotation of the audio was performed by LDC following the VAST Speech Activity Detection Guidelines (Version 1.7 – May 15, 2015 by LDC) [5].

The VAST data set presented several challenges for speech-analytic systems including audio compression, multiple languages, diverse topics, diverse recording equipment and quality, diverse background sounds, high-decibel background noise, low-decibel background babble, overlapping speech, and diverse real-world environments.

3.3 Babel Data

Low Resource Language

An unreleased subset was drawn from previously exposed Pashto language that was developed by Appen for the low-resource language (LRL) Babel program for IARPA (Intelligence Advanced Research Projects Activity) [6]. The IARPA Babel program focused on underserved languages and sought to develop an innovative, speech-recognition technology to be rapidly applied to any human language to improve keyword-search performance over substantial amounts of recorded speech.

While the Pashto language data for the system development phase in the Pilot OpenSAT was previously exposed, the Pashto language data for the system evaluation phase was not. It was a newly exposed subset of the collection created specifically for the Pilot OpenSAT evaluation. The goal, in part, was to leverage the previous effort developed for the IARPA Babel program by providing developers another system-evaluation opportunity for this language.

The Pashto audio collection presented challenges for speech-analytic systems such as low-resource, foreign language in telephone conversations, multiple microphone types (mobile and landline), and real-world environment (street, home or office, public place, inside vehicle). The primary challenge is the low resource language because, by definition, there are few resources available for systems development and training.

4 Participants

Overall, 33 sites (teams) with 75 researchers from 20 countries registered for the OpenSAT evaluation, of those, 22 teams with 43 researchers submitted systems results for evaluation. Teams that submitted systems output are listed below with their associated organization, country, and number of team members.

Table 2. Table shows tasks performed and the data domains used by each team

Team ID	Members	Organization	Country	SSSF			VAST	Babel		
				SAD	ASR	KWS	SAD	SAD	ASR	KWS
Audias-ATVS	1	Universidad Autonoma de Madrid	Spain	x			x	x		
BoUn_Team	1	Bogazici Universitesi	Turkey							x
BUT-Speech	1	Brno University of Technology	Czech Republic	x	x				x	
CMU	2	Carnegie Mellon University	USA	x			x	x	x	
CPqD	3	Centro de Pesquisa e Desenvolvimento em Telecomunicações	Brazil				x	x		
CRIM	3	Computer Research Institute of Montréal	Canada	x	x		x	x	x	
CRSS	3	The University of Texas at Dallas	USA	x						
Elektronika	2	Šiauliai University	Lithuania	x			x	x		
I2R	1	Institute for Infocomm Research	Singapore	x	x		x	x	x	
IDIAP-Team	1	Istituto Dalle Molle di Intelligenza Artificiale Percettiva	Switzerland	x						
IIT-Guwahati	3	Indian Institute of Technology Guwahati	India	x	x	x	x	x	x	x
IntelligentVoice	3	Intelligent Voice Limited	England	x			x		x	
JHU-CLSP	4	Johns Hopkins University	USA				x	x	x	x
LaBRI-Speech	1	Laboratoire Bordelais de Recherche en Informatique	France	x			x	x		
MIRACL	1	Not available	Not available					x		
QUT-Team	1	Queensland University of Technology	Australia	x			x	x		
RUN	1	Radboud University	Netherlands				x			

SRI+QUT	3	SRI International + Queensland Univ. of Technology	USA+ Australia	x			x			
SRI+RUN	2	SRI International + Radboud University	USA+ Netherlands				x			
SRI-STAR	1	SRI International	USA				x			
THUEE	3	Tsinghua University, Beijing	China	x			x	x	x	x
TRIst	2	MEF University	Turkey					x	x	x

5 Results

5.1 SAD Results

The plots in Figures 2, 3, and 4 show the system-output scores by data domains for teams that submitted them[5]. Teams with low DCF scores performed better in speech-activity detection than teams with high DCF scores. Factors such as background babble or foreground noise can cause both false positives and false negatives. Data that combines stressed speaking and radio- channel noise with strong background noise, such as the Sofa Super Store fire data, can be challenging to analyze. Data that contains such diverse acoustics as those in the VAST dataset can also pose varying challenges throughout the data set. System-output scores are plotted by their DCF value - low to high, left to right - with respective false-negative and false-positive probabilities.

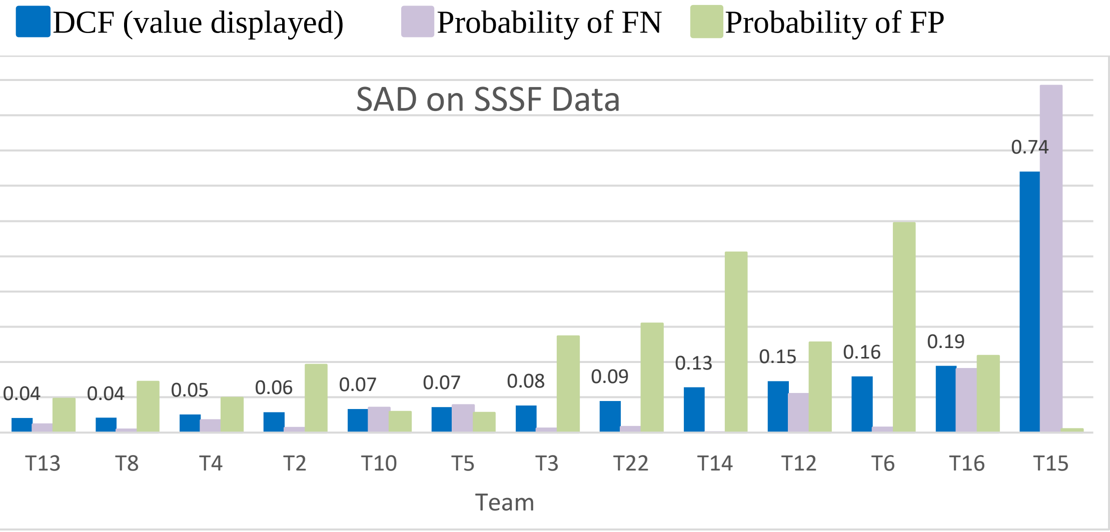


Figure 2. SSSF Data, SAD Results

[5] Note, some teams did not participate in all three data domains

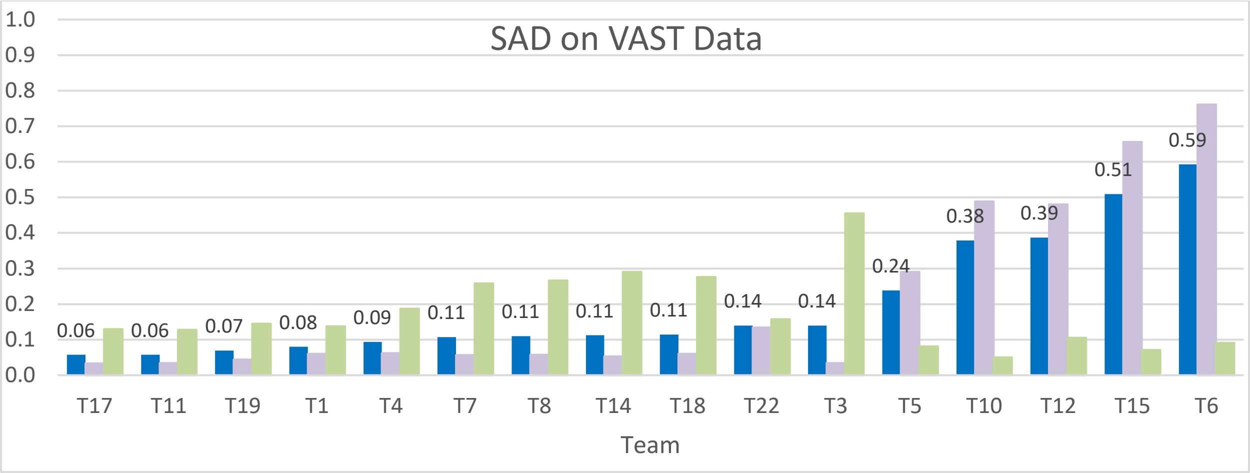


Figure 3. VAST Data, SAD Results

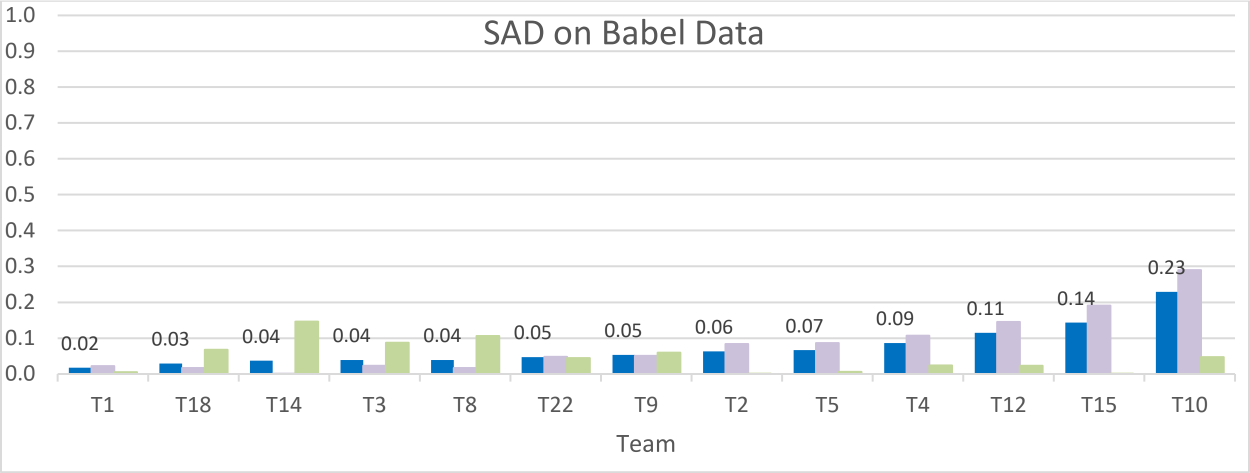


Figure 4. Babel Data, SAD Results

Teams that participated in all three data domains for SAD

The plot in Figure 5 displays a contrast of DCF scores for all three data domains by team, where they exist. The scores are grouped by teams and ordered low-to-high, left-to-right using the Babel scores. The Babel data provided the lowest (best) DCF scores for seven of the ten teams; none in VAST, and three in SSSF. The VAST data provided the highest (worst) DCF scores in eight of the ten teams, and two in SSSF.

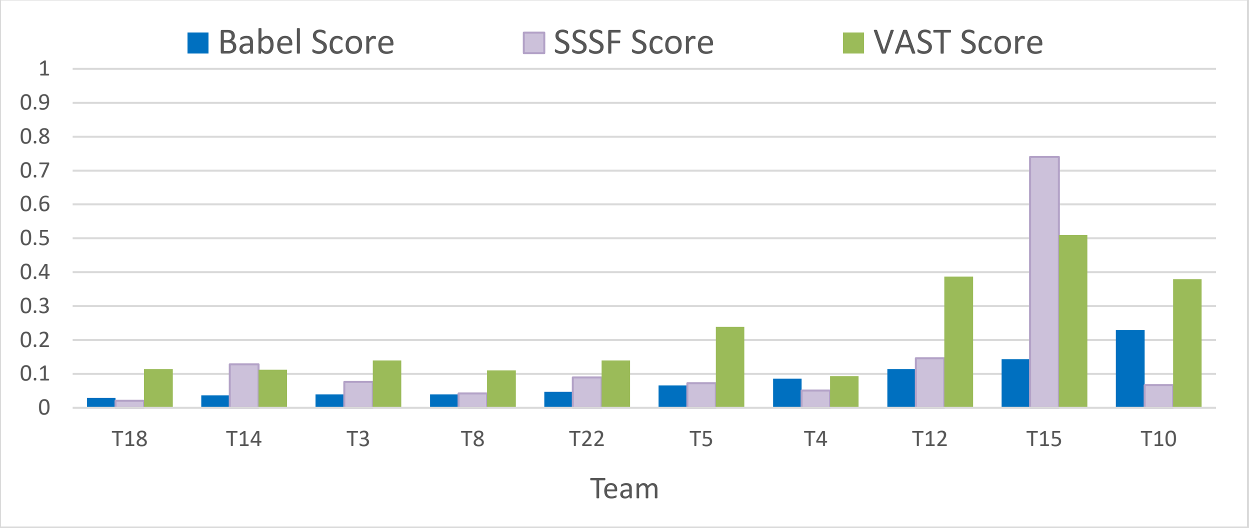


Figure 5. Contrast of DCF scores for SAD across the three data domains
Y axis = Detection Cost Function (DCF) Values for Speech Activity Detection (SAD)

5.2 KWS Results

The KWS task was performed on the Babel and SSSF datasets only. Plots show ATWV values high to low, left to right. Higher ATWV scores are betters scores.

Comments from the post evaluation teleconference indicated that KWS results during the development phase with the SSSF data were bad enough to discourage continued development for that task. The SSSF dataset was apparently extremely challenging for the KWS task. The limited amount of SSSF development data available relative to the evaluation data would have contributed to making this task more challenging.

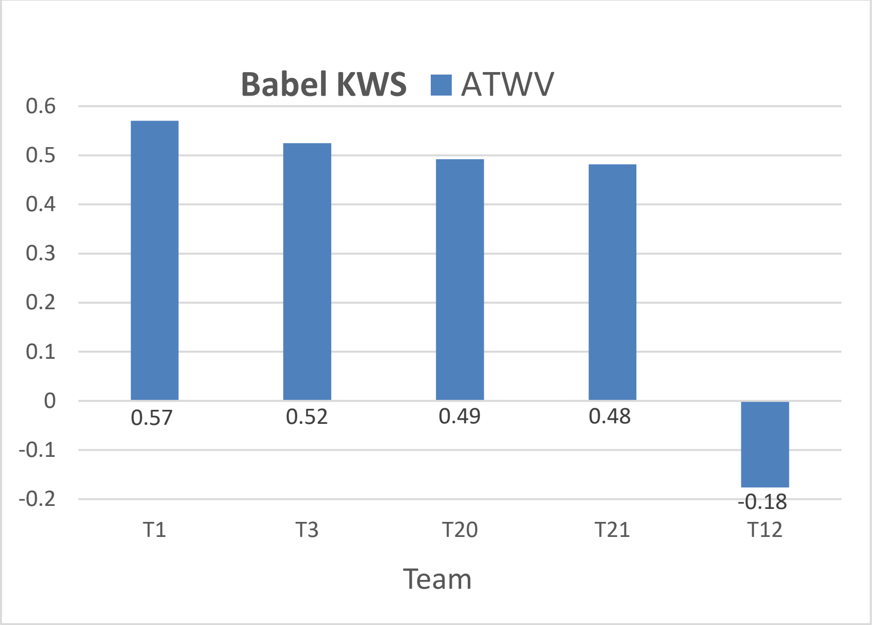


Figure 6. Babel Data, KWS Results

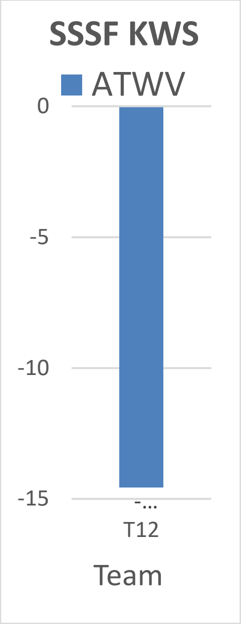
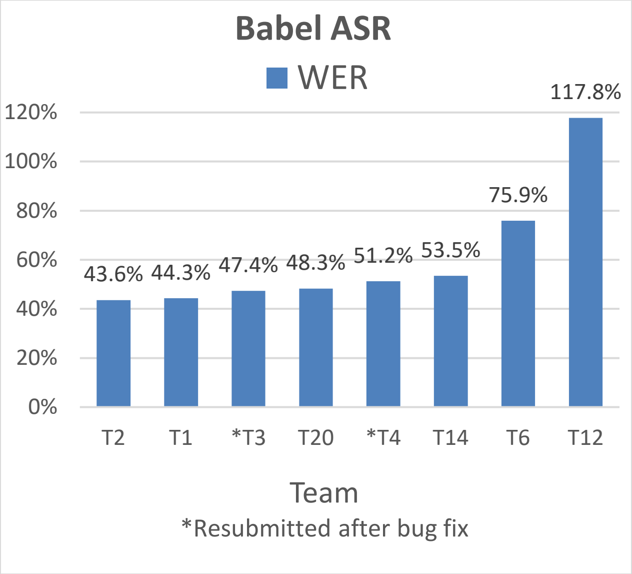


Figure 7. SSSF Data, KWS Results

5.3 ASR Results

The ASR task was performed on the Babel and SSSF datasets only. Plots show WER values low to high, left to right. Lower scores are betters scores. Again, the limited amount of SSSF development data available relative to the evaluation data would contrubute to making this task challenging.



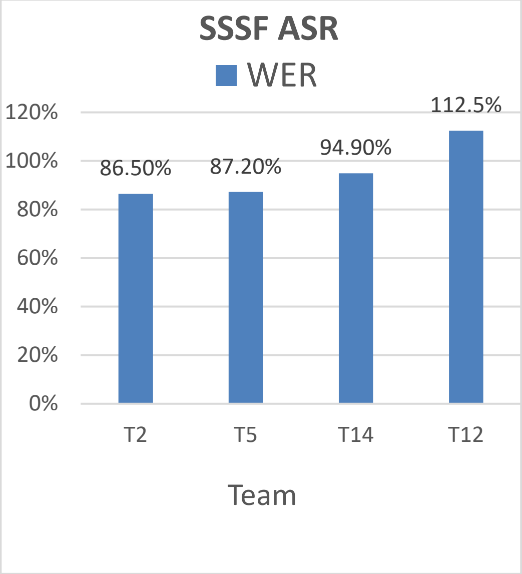


Figure 8. Babel Data, ASR Results

Figure 9. SSSF Data, ASR Results

6 Analysis and Review

6.1 SSSF Dataset Review

The total audio time was fixed. To provide sufficient audio time for the evaluation phase, a less than desirable length of audio time was available for the development phase. This created an imbalance in the amount of development data relative to the amount of evaluation data. Ideally, much more development data than was made available is preferred by developers to optimally train/develop their systems. As learned from the post evaluation teleconference, the shorter audio time of reference data made available for the development phase was not sufficient for many developers to tune their systems adequately. This caused some developers to discontinue development efforts; they did not submit system output for scoring, particularly for KWS and ASR tasks, because results were so poor during the development phase.

Another lesson learned from the post evaluation teleconference is that while transcription is preferred for all development data, additional development data is preferred even if not transcribed, since it is still useful for system training and development, particularly if it is available from the same dataset.

6.2 VAST Dataset Review

During the post OpenSAT Pilot teleconference, input from participants included a concern over the accuracy of the SAD annotation (reference material) on the VAST data. The VAST data domain may have been the most challenging of the three for systems analytics; or, as noted from post evaluation comments, the reference annotation may have had some issues.

On follow up, examples of questionable annotation in the VAST dataset were provide to NIST for review. Two methods were used to review and confirm whether there were errors in the annotation and, if any, measure the level of discrepancy. The first method applied solely to the examples provided.

The first method involved reviewing the examples of discrepancies submitted by one of the participants. Six examples of questionable annotations from the development set were submitted. NIST performed a manual annotation of those examples and contrasted it against the VAST dataset reference annotation.

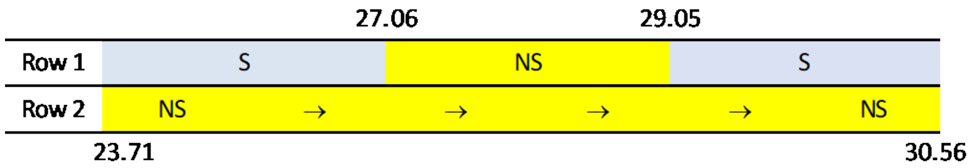
A second method looked at seven teams that had achieved the best seven overall DCF scores from the evaluation data and reviewed their worst ten file scores to see if there were common files among those seven teams and if there were similarities among those files.

For the following examples, values are time (in seconds) from the beginning of the audio file. S and NS denote speech and non-speech, respectively.

6.2.1 Files from the VAST development set reviewed – First method.

Example 1: VVC011246 audio file for the development phase

In this example, the participant indicated that there is “no audible speech” beginning at 23.71 s and continuing to the end (30.56 s) of these segments, including the segments shown in the reference annotation as speech (S).



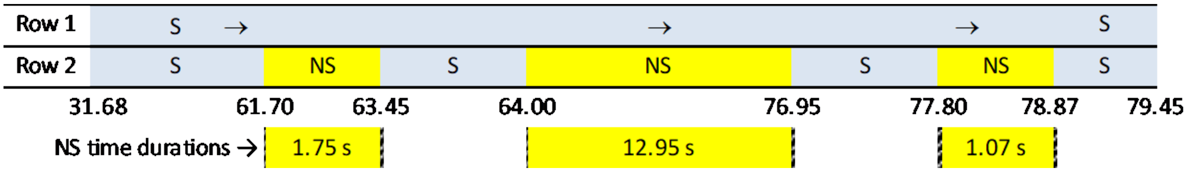
Row 1: Reference annotation
Row 2: Manual annotation of area of interest (23.71 s through 30.56 s)

Figure 10. VVC011246 audio file

On review, the “background speech” in the two speech segments in the reference annotation is barely audible. As a result, the two speech segments may be questionable for certainty that it is speech by manual annotation. It may also be sufficiently difficult or unrealistic to expect a system to detect speech, if speech actually is present.

Example 2: VVC031565 audio file for the development phase.

The participant indicated that there is “no audible speech” from 63.45 to 79.45 of these segments, including the speech segment shown in the reference annotation. The reference annotation shows continuous speech from 31.68 s to the end of the file, 100.70 s. The Example 2 diagram shows only the time period between 31.68 s through 79.45 s to focus on the participant’s time-span in the example that was submitted.



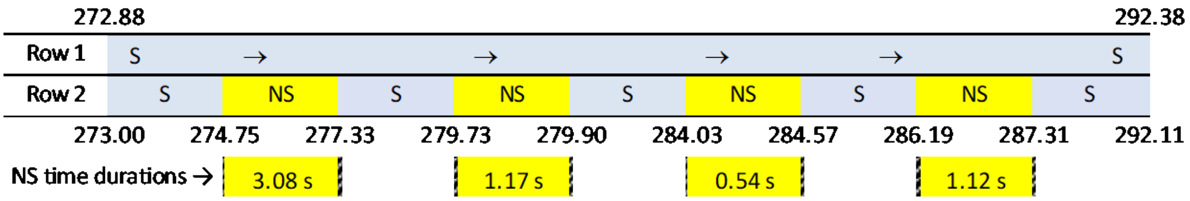
Row 1: Reference annotation
Row 2: Review of participant’s area of concern (61.60 s through 79.45 s

Figure 11. VVC031565 audio file

On review, a manual inspection of the 31.68 s - 100.70 s time period showed eleven non-speech segments greater than 1.00 s in duration plus three less than 1.00 s where the reference annotation shows all speech. In the time-span of the example submitted as shown in example 2, there are clearly non-speech segments, including a 12.95 s time-span, where the reference annotation shows all speech.

Example 3: VVC004451 audio file for the development phase.

The participant indicated that there is “no audible speech” in three areas within the time span of 272.88 s through 292.39 s, where the reference annotation indicates all speech.



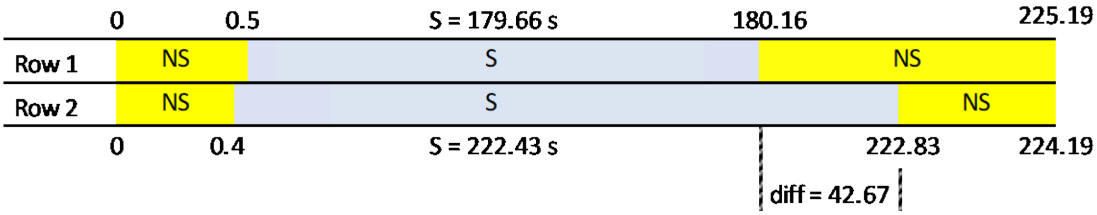
Row 1: Reference annotation durations
Row 2: Review of participant’s area of concern (273.00 s through 292.11 s)

Figure 12. VVC004451 audio file

The review shows where there are four non-speech segments found by manual annotation, three of the four are greater than one second, compared to continuous speech in the same region for the reference annotation.

Example 4: VVC026043 audio file for the development phase.

This example indicated that there is speech in an area where the annotation shows non-speech. The area in question is in non-speech segment in the reference annotation directly after the annotated speech ends at 180.16 s. There is a 42.67 s discrepancy between the speech and non-speech segments, comparing the reference annotation with the review of the participant’s area of concern.

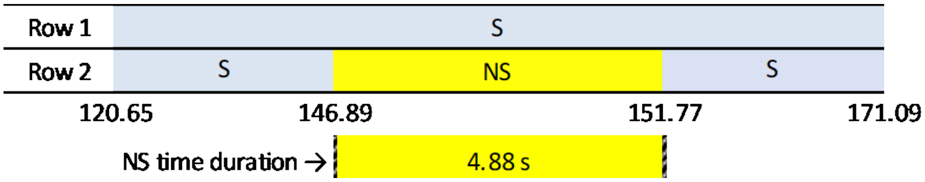


Row 1: Reference annotation
Row 2: Review of participant’s area of concern (180.16 s through 222.83 s)

Figure 13. VVC026043 audio file

Example 5: VVC009379 audio file for the development phase.

This example indicated that there is “no audible speech” from 146.89 to 151.77. The reference annotation shows continuous speech from 120.65 s through 171.09 s. On review, there is no speech between 146.89 s and 151.77 s (4.88 seconds of non-speech).

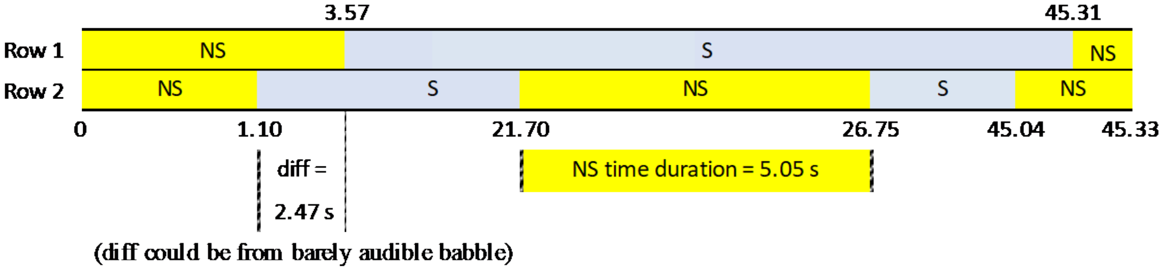


Row 1: Reference annotation
Row 2: Review of participant's area of concern (146.89 s through 151.77 s)

Figure 14. VVC009379 audio file

Example 6: VVC041917 audio file for the development phase.

This example indicated that there is "no audible speech" from 21.70 to 26.75. The reference annotation shows continuous speech from 3.57 s through 45.31 s. In review by manual annotation, there is no speech between 21.70 s and 26.75 s (5.05 seconds of non-speech).



Row 1: Reference annotation
Row 2: Review of participant's area of concern (273.00 s through 292.11 s)

Figure 15. VVC041917 audio file

6.2.2 Files from the VAST evaluation set reviewed – Second method

The DCF score returned to the participant at the conclusion of the evaluation is the average of the 200 DCF scores representing all the files in the evaluation dataset. In this step, we only looked at the ten worst scored files out the 200 files for each of the best performing teams, i.e., seven teams with the best overall averaged DCF scores.

Table 3 shows the seven teams with their worst ten scores, with the worst being at the top of the list for each team. Files are colored to show common files. Several files are common among the worst scored files among the teams

Table 3. Ten worst scores of the seven best scoring teams

	T11	T17	T19	T1	T4	T7	T8
1	VVC010551	VVC010551	VVC015475	VVC027464	VVC015258	VVC016338	VVC015475
2	VVC037067	VVC037067	VVC010551	VVC033478	VVC015475	VVC027464	VVC009090
3	VVC007792	VVC007792	VVC016338	VVC010665	VVC016338	VVC015475	VVC010551
4	VVC009090	VVC009090	VVC024151	VVC008286	VVC008947	VVC010551	VVC033624
5	VVC033431	VVC033431	VVC037067	VVC036166	VVC010551	VVC024151	VVC000427
6	VVC000427	VVC024151	VVC036371	VVC010551	VVC002875	VVC000512	VVC037067
7	VVC024151	VVC013898	VVC013898	VVC036371	VVC012193	VVC011947	VVC007792
8	VVC013898	VVC000427	VVC007792	VVC009090	VVC013898	VVC013898	VVC024151
9	VVC016539	VVC016338	VVC009090	VVC026328	VVC036166	VVC037067	VVC036166
10	VVC029648	VVC016539	VVC007438	VVC024151	VVC024151	VVC032541	VVC027464

Table 4. VVC010551 audio file review

File # VVC010551 (teams T11, T17, T19, T1, T4, T7, T8)

	DCF value Score	system Miss FN Sum	system FA FP Sum	system correct TN Sum	system correct TP Sum	reference Speech time Sum	reference Non-Speech time Sum
Team 11		32.653	188.487	2.902	186.346	218.999	191.389
Team 17		29.313	188.477	2.912	189.686	218.999	191.389
Team 19		32.082	188.437	2.952	186.917	218.999	191.389
Team 1		12.584	190.714	0.675	206.415	218.999	191.389
Team 4		32.012	188.455	2.934	186.987	218.999	191.389
Team 7		19.252	189.013	2.376	199.747	218.999	191.389
Team 8		18.254	180.129	11.26	200.745	218.999	191.389
	Results by manual annotation →					374.420	73.830

7 out of the 7 teams with the best DCF scores had this file in their worse 10 scores.

The reference non-speech time is significantly higher than found in a manual annotation review. The high non-speech time in the reference would result in a high FP performance by a system.

Comments/observations:
English language.
Sounds like a teacher with a group of young children, perhaps preschoolers.
Lots of singing, clapping.
Music in the background - not part of the singing.

Table 5. VVC024151 audio file review

File # VVC024151 (teamsT11, T17, T19, T1, T4, T7, T8)

	(system Miss) FN Sum	(system FA) FP Sum	(system correct) TN Sum	(system correct) TP Sum	(reference) Speech time Sum	(reference) Non-Speech time Sum
Team 11	46.525	0.290	9.096	103.010	149.535	9.386
Team 17	48.131	0.320	9.066	101.404	149.535	9.386
Team 19	56.555	0.170	9.216	92.980	149.535	9.386
Team 1	43.337	1.567	7.819	106.198	149.535	9.386
Team 4	57.208	0.140	9.246	92.327	149.535	9.386
Team 7	61.355	0.010	9.376	88.180	149.535	9.386
Team 8	37.804	2.533	6.853	111.731	149.535	9.386
	Results by manual annotation →				65.91	99.46

7 out of the 7 teams with the best DCF scores had this file in their worst 10 scores.

The reference speech time is significantly higher than found in a manual annotation review, and the non-speech time is low. The higher than expected speech time in the reference would result in a high FN performance by a system.

Comments:
Foreign language.
Single speaker.
Varying background noise, sometimes loud.
Echo in some of the background noise.
Conversations in distance background noise, barely audible. There could be uncertainty in annotation guideline-understanding and/or inconsistency in annotator speech detection versus a system's speech detection.

Table 6. VVC009090 audio file review

File # VVC009090 (teams T11, T17, T19, T1, T8)

	(system Miss) FN Sum	(system FA) FP Sum	(system correct) TN Sum	(system correct) TP Sum	(reference) Speech time Sum	(reference) Non-Speech time Sum
Team 11	1.690	1.14	0	250.539	252.229	1.14
Team 17	2.450	1.14	0	249.779	252.229	1.14
Team 19	1.470	1.14	0	250.759	252.229	1.14
Team 1	5.366	1.14	0	246.863	252.229	1.14
Team 4	4.810	0.54	0.6	247.419	252.229	1.14
Team 7	0.130	1.14	0	252.099	252.229	1.14
Team 8	19.536	1.14	0	232.693	252.229	1.14

5 out of the 7 teams with the best DCF scores had this file in their worst 10 scores.

The reference non-speech and speech times may be about right, and this audio file might just be a significant speech analytic challenge for systems. The non-speech total time is very small relative to the total speech time, causing an extreme speech to non-speech ratio of 221.25 to 1, respectively. An extreme ratio can cause a high DCF score when a small system error occurs on the smaller time.

Comments:
Foreign language.
Some coughing, laughing, low rough scratchy voice in places.
No background noise for almost all of the audio.
Occasionally people talking behind others.
Several people in the conversation. Some overlap.

Table 7. VVC013898 audio file review

File # VVC013898 (teams T11, T17, T19, T4, T7)

	(system Miss) FN Sum	(system FA) FP Sum	(system correct) TN Sum	(system correct) TP Sum	(reference) Speech time Sum	(reference) Non-Speech time Sum
Team 11	40.058	0	7.575	86.710	126.768	7.575
Team 17	41.858	0	7.575	84.910	126.768	7.575
Team 19	44.308	0	7.575	82.460	126.768	7.575
Team 1	33.658	0	7.575	93.110	126.768	7.575
Team 4	52.388	0	7.575	74.380	126.768	7.575
Team 7	47.228	0	7.575	79.540	126.768	7.575
Team 8	29.890	2.045	5.530	96.878	126.768	7.575

5 out of the 7 teams with the best DCF scores had this file in their worst 10 scores.

The reference non-speech time appears that it may be lower than the actual non-speech time because of consistent high FN sums across systems.
Comments:
Foreign language.
Noisy echo kind-of background like in a large open facility, e.g., shopping mall or auditorium.
The background noise decibel level near the voice decibel level.

Table 8. VVC037067 audio file review

File # VVC037067 (T11, T17, T19, T7, T8)

	(system Miss) FN Sum	(system FA) FP Sum	(system correct) TN Sum	(system correct) TP Sum	(reference) Speech time Sum	(reference) Non-Speech time Sum
Team 11	9.700	60.286	0.560	255.208	264.908	60.846
Team 17	7.740	59.186	1.660	257.168	264.908	60.846
Team 19	14.355	58.286	2.560	250.553	264.908	60.846
Team 1	0.034	60.846	0.000	264.873	264.908	60.846
Team 4	25.798	49.937	10.909	239.110	264.908	60.846
Team 7	20.771	53.552	7.294	244.137	264.908	60.846
Team 8	18.061	51.976	8.870	246.847	264.908	60.846

5 out of the 7 teams with the best DCF scores had this file in their worse 10 scores.

The reference non-speech time appears that it may be higher than actual non-speech time because of consistent high FP sums across systems. After listening to the complete audio file, it appears this observation would be supported by manual annotation.

Comments:
Foreign language.
Very soft-spoken in the beginning.
Some singing, noisy background relative to voice in some places, a little laughing.

File #s VVC007792, VVC015475, and VVC016338

4 out of the 7 teams with the best DCF scores had these three files in their worse 10 scores. In each file, the reference non-speech and speech times may be about right. Also, in each file, the total non-speech time is very small at 0.1 s, 0.31 s, and 0.00 s respectively, and appears correct. The significantly low non-speech time can impact an individual file's DCF score significantly, because even a small amount of error will create a larger error rate. Audio files with such an extreme imbalance between speech and non-speech times, especially where one approaches zero, for example 99% speech - 1% non-speech, will impact error probability rates significantly from a minor non-speech error, and thus the DCF value for that individual file.

The remaining two of the ten worst files had only three or fewer teams with those in their bottom ten scores. These files were not reviewed.

6.3 Babel Dataset Review

The Babel data provided the best scores for all tasks, and there were no issues commented on during the post evaluation teleconference. One advantage developers may have had with this data domain is that the development set had previously been made available in the IARPA Low Resource Language program. Many of the developers could have already been familiar with this data domain and may also have been able to piggyback on previous development work.

7 Summary

The Pilot OpenSAT Evaluation was successful in 1) attracting a large number of speech-analytic-system developers for different tasks and 2) exchanging comments and suggestions during the post evaluation teleconference. The post-evaluation teleconference provided valuable insights into challenges the developers faced and lessons learned for improving the next evaluation. Recommendations for the next evaluation include 1) balance the overall speech to non-speech ratio in the audio dataset closer to the 50/50 direction, 2) provide much more data for system development, this can include other data not transcribed from the same domain, and 3) check/review reference annotations for accuracy.

The SAD task was especially valuable for two reasons. First, it is a prerequisite for performing many other speech-analytic tasks, Second, it was particularly useful as a prelude to an NSA speaker/language recognition project with the NSA VAST data collection. This task also brought other developers into the evaluation with an opportunity to try other tasks and data domains. With a larger dataset for public-safety communications in the next evaluation, particularly for systems development, we anticipate an increased participation for that data domain.

While most of the teams participated in the SAD task, several also participated in the KWS and ASR tasks. The three different data domains provided several combinations of tasks and datasets from which developers could choose. This alone has accomplished the goal of providing the potential for information sharing among developers, who that historically do not collaborate . A greater than expected number of teams participated in the evaluation and, based on teleconference comments, developers are anticipating the next OpenSAT evaluation.

The speech in the SSSF dataset was often difficult to understand, making it particularly challenging for the KWS and ASR tasks as these require accurate transcription output from speech analytic systems to achieve good performance scores.

The VAST dataset was available for the SAD task only. Based on participants' scores it appeared to be more challenging, generally, than the Babel and SSSF datasets. However, participants provided comments suggesting that the annotation may have had inaccuracies. Several examples were provided to NIST for a follow up and a review. Indeed, NIST confirmed that there were some inaccuracies in the annotation for the reference key for those examples. Since some of those inaccuracies were significant, one would expect the files containing those inaccuracies to result in poorer DCF values. Additionally, a margin of negative impact on overall systems performance scores would be expected as well.

To better understand the probable causes for the VAST dataset to appear challenging, NIST conducted a further review and analysis of the 200 files in the dataset. One potential cause is already known, questionable annotation. NIST staff were looking for additional causes. The staff found extreme differences between speech and non-speech times in individual files. For example, the staff found that 55 of the 200 reference files had no (0) non-speech times using the 0.5 s collar. Filtering the 200 files for those that have a speech to non-speech ratio of 70-30 percent or 50-50 percent, again using the 0.5 s collar, only 23 files would remain for DCF calculation. Over half of the 200 files have a less than 10% non-speech, i.e., less than a 90-10 ratio. Even if there were no collar used (0.0 s collar), there would still only be 37 files left after filtering for a ratio of 70-30 speech-to-non-speech.

Also, the method for calculating the DCF value to indicate system performance was made by averaging the sum of all the DCF values from the 200 files in the dataset. An alternate method would be to first separately sum the individual speech and non-speech segment times for that dataset and then calculate the DCF value. This would be done for both the reference data and the system's output. The alternative method may minimize potential biasing in calculating a system's overall DCF value from "lopsided" speech-to-non-speech-ratio files using the averaging method. There were many lopsided S-NS files in the VAST dataset. By comparing these two methods for several teams on the VAST dataset, the resulting DCF values between the two methods were significantly different. The current DCF averaging method shows DCF values indicating better system performance than when using the alternative, time-sum-first method.

With 1) the possible negative impact on DCF values, which depend on the scale-of-reference annotation inaccuracies, and 2) the positive impact on DCF values from "lopsided" files in DCF averaging, it is uncertain how accurate the DCF values reflect actual level of system performance from the VAST data. However, the ranking between the systems is expected to be reasonably accurate since these factors should have the same negative/positive impact on all systems. A further review for confirmation of this expectation would be needed.

The DCF values were close to the same between the two methods for the Babel and SSSF datasets. Even though the DCF values were close to the same for these datasets, the alternative "time summing" method rather than the current DCF averaging method is recommended for a system's overall performance and ranking for the next evaluation. It is also recommended that the DCF values for individual files are still calculated, averaged, and made available to provide both a comparison and the option for a more in-depth analysis if desired.

8 References

[1] Office of the Director of National intelligence (website)
https://www.iarpa.gov/index.php/research-programs/babel

[2] Evaluation Plan for the NIST Open Evaluation of Speech Activity Detection (OpenSAD15), version 10.0, May 12, 2016
https://www.nist.gov/sites/default/files/documents/itl/iad/mig/Open_SAD_Eval_Plan_v10.pdf

[3] DRAFT KWS16 Keyword Search Evaluation Plan
https://www.nist.gov/sites/default/files/documents/itl/iad/mig/KWS16-evalplan-v04.pdf

[4] Charleston Sofa Super Store Phase II report, May 15, 2008)
http://downloads.pennnet.com/fe/misc/20080515charlestonreport.pdf

[5] VAST: Video Annotation for Speech Technologies
Corpus Specification, LDC version 0.7, Oct. 31, 2013

[6] IARPA Babel Data Specification for Performers
https://www.nist.gov/sites/default/files/documents/itl/iad/mig/IARPA_Babel_Specification-02062013.pdf August 26,2013