The SAFE-T Corpus: A New Resource for Simulated Public Safety Communications

Dana Delgado, Kevin Walker, Stephanie Strassel, Karen Jones, Christopher Caruso, David Graff
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street Suite 810, Philadelphia, PA, 19104, USA
{foredana, walkerk, strassel, karj, carusocr, graff}@ldc.upenn.edu

Abstract

We introduce a new resource, the SAFE-T (Speech Analysis for Emergency Response Technology) Corpus, designed to simulate first-responder communications by inducing high vocal effort and urgent speech with situational background noise in a game-based collection protocol. Linguistic Data Consortium developed the SAFE-T Corpus to support the NIST (National Institute of Standards and Technology) OpenSAT (Speech Analytic Technologies) evaluation series, whose goal is to advance speech analytic technologies including automatic speech recognition, speech activity detection and keyword search in multiple domains including simulated public safety communications data. The corpus comprises over 300 hours of audio from 115 unique speakers engaged in a collaborative problem-solving activity representative of public safety communications in terms of speech content, noise types and noise levels. Portions of the corpus have been used in the OpenSAT 2019 evaluation and the full corpus will be published in the LDC catalog. We describe the design and implementation of the SAFE-T Corpus collection, discuss the approach of capturing spontaneous speech from study participants through game-based speech collection, and report on the collection results including several challenges associated with the collection.

Keywords: collection, speech resources, game protocol, speech analytic technologies, transcription

1. Introduction

We introduce a new resource, the SAFE-T (Speech Analysis for Emergency Response Technology) Corpus, designed to address the need for training, development and test data representing public safety communications. The SAFE-T Corpus was developed by Linguistic Data Consortium to support the National Institute of Standards and Technology OpenSAT (Speech Analytic Technologies) Evaluation campaign, and was first used as part of the OpenSAT 2019 evaluation. The goal of OpenSAT is to advance speech analytics including speech activity detection (SAD), keyword spotting (KWS) and automatic speech recognition (ASR), across multiple data domains (NIST, 2019).

One especially challenging domain for speech analytic technologies is public safety communications, which are characterized by prominent background noise, radio channel noise[1], speech under stress and urgent speech, the Lombard effect, and other properties. To address the need for test data in the public safety domain, LDC designed the SAFE-T (Speech Analysis for Emergency Response Technology) Corpus to collect data from speakers engaged in collaborative problem-solving activities that would be representative of public safety communications in terms of speech content, noise types and noise levels. To support OpenSAT goals, the corpus design required collection from 100 speakers recruited from the North American English speaking population, with each speaker providing a minimum of 2 hours of speech recordings, and a total audio data volume of 291 hours.

The collection sought to elicit speech exhibiting specific features found in public safety communications. These include the Lombard effect in which speech behavior is altered due to presence of prominent background noise, a range of high and low vocal effort, speaker stress due to the perception of situational urgency, spontaneous speech, and lexical items that occur in the public safety domain. A game-based collection protocol was used to elicit spontaneous collaborative speech from recruited participants, and naturally occurring audio from real world emergency events was used as background noise played at two distinct noise levels to produce varying levels of vocal effort, Lombard effect and urgent speech from the recruited speakers.

Each recording session consisted of two thirty-minute games of Flash Point Fire Rescue (Lanzing, 2011), a cooperative board game in which two players have to work together to rescue victims from a burning house. The game elicits natural conversation with vocabulary relevant to the intended domain. Stress and urgency build as the game proceeds with additional pressure deliberately introduced by adding time limits on completing game tasks. During recording sessions, each player wore a headset with a built-in microphone through which they heard not only their game partner's speech but also a variety of emergency event noises at different volume levels. Each player's speech was recorded to a separate channel and mixed with the background noise recordings to create training, development and test data, portions of which were manually transcribed.

A portion of the resulting SAFE-T Corpus audio recordings, metadata and transcripts were selected for use in OpenSAT 2019, while additional data has been held back for use in future OpenSAT evaluations. After its use in OpenSAT, all SAFE-T corpus data will be published in the LDC Catalog, making it available to the research community at large.

[1] The SAFE-T corpus was originally designed to include retransmission of collected speech recordings over various radio channels, but this was dropped in favor of additional transcription to support the needs of OpenSAT evaluation.

2. Prior Work

There have been a number of previous collection efforts that also focused on game-based speech and/or high vocal effort.

The English-L2 Child Learner Speech Corpus, for example, collected by the Université de Genève FTI / TIM (Baur, Rayner & Tsourakis, 2014) used web-based gamification to record the speech of German speaking students learning English. The SAFE-T collection methods were similar to that of the Columbia Games Corpus (Gravano, Hirschberg, 2011) which was a speech collection in which the speakers sit across from one another separated by a physical barrier and only communicate by voice while engaged in a cooperative game. The Speech in Noisy Environments (SPINE) Training Audio Corpus (Schmidt-Nielsen, et al., 2000) that was developed for the Department of Defense (DoD) Digital Voice Processing Consortium (DDVPC) by Arcon Corp. and distributed by LDC is perhaps the most similar collection. SPINE combined both collaborative game-based speech collection and transcribing speech in noisy military environments.

As with these prior efforts, SAFE-T used game play to elicit speech from recorded speakers and used background noise to elicit high vocal effort, but instead of military environments, the SAFE-T corpus collection sought to mimic public safety communications. Also in contrast to prior collections, SAFE-T focused on eliciting a wide range of vocal effort, both high and low, as well as obtaining highly cooperative speech from speakers in close proximity, working together to solve domain-relevant problems.

3. Data Requirements

To support the requirements of OpenSAT evaluations for 2019 and beyond, the SAFE-T Corpus needed to include a minimum of 291 hours of audio from at least 100 unique speakers, with multiple 30-minute recordings per speaker, and at least 122 hours of the collected audio manually transcribed. Each recording included two background noise types and two noise levels, where noise types are the kinds of background noise heard by participants and noise levels are the decibel ranges used for the background noise during a particular section of the recording. The collection protocol was designed to elicit a ratio of 40% speech and 60% non-speech on average, and to yield one to three minutes of urgent speech per 30-minute recording.

4. Collection Protocol

Each recording session consisted of two subjects who knew each other playing a collaborative problem solving domain-relevant board game. Participants were separated by a physical barrier and wore headphones through which they heard background noise of varying types and levels. Their speech was recorded via a high quality head-mounted microphone, with a separate channel for each speaker. Recording sessions also included a game master who provided instructions and managed the recording session, and a technical assistant who set up the recording equipment and managed the files and metadata. Each recording session consisted of two 30-minute recordings, and speakers participated in a minimum of 2 and a maximum of 4 recording sessions, with no more than one session per person per day.

After a recording session concluded, the recorded speech from each speaker was mixed with the background noise recordings heard by that speaker to produce the training, development and evaluation data used for OpenSAT. The unmixed, clean channel recordings were used to produce manual reference transcripts used for OpenSAT system training, development and testing.

4.1 Game Requirements

LDC researched multi-player board games to identify those that involved collaborative problem solving, elicited natural and spontaneous speech, resulted in a significant quantity of speech from each recorded speaker, required a high degree of interaction between speakers, produced domain-relevant vocabulary, and yielded levels of vocal intensity commonly found in operational speech, including both high and low vocal effort. It was also necessary for the board game to be relatively fast and easy for participants to learn as well as enjoyable for them to play. The game's duration also needed to be appropriate for a reasonable length recording session. Online games were not compatible with the collection protocol for several reasons, primarily logistical. The collection protocol required direct interactions between speakers, with players physically present in the same room; this setup lends itself more naturally to board games as opposed to online games. In addition, the kinds of sound effects found in many online games would introduce noise types that were outside the scope of the collection targets. Finally, it was necessary to select a game whose rules could be easily manipulated in order to induce more speech of the type required for the corpus; online games were less amenable to such modifications compared to board games.

4.1.1 Game Selection

After testing multiple candidate games, we selected the board game Flash Point Fire Rescue. The game involves 2-4 participants working together in order to rescue people and animals from a burning building before it collapses. The game requires intensive collaboration to solve domain-relevant problems: players act as firefighters extinguishing smoke and fire while moving through a burning house to check on and rescue points of interest. The game has a reasonably short setup time and learning curve for first time players; it has a duration compatible with a reasonable length recording session; and it tends to prompt communication from all players, who must become increasingly cooperative as the game progresses. The premise of the game also elicits domain-related vocabulary and urgent speech, which can be further increased through rule modifications and additions.

Although the game allows for 2-4 participants, we found during testing that games involving only 2 players were optimal for our goals. Using only 2 players elicited a sufficient amount of speech from each person without

either one tending to dominate, while using 4 players did not result in sufficient speech and did lead to dominant speakers. Therefore, we used 2 player games for all recording sessions.

4.1.2 Game Modification

To achieve various recording goals including quantity and urgency of speech and presence of in-domain vocabulary, and to control the duration of the game we made several changes to the rules, including the following:

- Players were asked to use two-way radio communication protocol, e.g. taking turns while speaking and using terms like "over", "roger", "copy" and "repeat", in order to produce more domain-relevant speech.
- Players were instructed to talk throughout their turns, narrating their actions and verbalizing their plans, in order to yield more speech per player.
- Game masters introduced a time limit for saving the next victim, in order to introduce more urgency in players' speech.
- Game masters requested regular status reports from players, including things like describing the location of the fire, status of each room of the house, location of victims and so on, in order to encourage players to produce more speech.
- Game masters issued a "radio failure" penalty for players who were not speaking enough; 3 radio failures resulted in the loss of a victim.

Beyond manipulating the game's rules to induce the desired speech content, quantity and quality from players, we also installed a physical barrier made of acoustic foam between the participants to maximize verbalization by preventing the use of eye contact and other non-verbal cues. The barrier also reduced the amount of interlocutor speech picked up by the speakers' microphones.

4.2 Participant Recruitment and Enrollment

Speakers were recruited in Philadelphia by word of mouth and by advertising to local emergency response organizations. Word of mouth was the most successful method of recruitment, partly because participants were instructed to bring their own friends for each game session. There were no hard requirements regarding speaker demographics, but there was a general goal to have the distribution reflect the first responder population (e.g. more male than female speakers).

An enrollment website provided information about the study to potential participants, explaining what would be involved in a typical recording session, privacy protections and participant compensation. Interested participants then enrolled in the study, providing basic demographic information including year of birth, city born/raised, education and sex. After enrollment was complete, an automated email was sent with the participant's assigned PIN and instructions to schedule their first recording session. The email also asked them to bring a friend along to the recording session to act as a game partner; the friend could either enroll in advance through the website or enroll onsite at the start of recording session. Participant pairings were allowed to be the same for every recording session, or they could vary from one session to the next.

All SAFE-T Corpus collection activities were conducted with review from the University of Pennsylvania's Institutional Review Board. All speakers in the corpus provided informed consent upon enrollment, and they were compensated for their participation.

4.3 Recording Sessions

4.3.1 Session Management

Each recording session lasted up to 90 minutes and was comprised of two 30-minute recorded games plus time for setup, a break and wrap-up.

Also present at each recording session were a game master and a technical assistant. The game master acted as a session manager, taking participants through the session from start to finish, while the technical assistant focused on ensuring the collection platform was operating as intended as well as handling the resulting audio recordings.

Game master responsibilities included:

- Check in the participants, enroll game partner if not pre-enrolled, offer refreshments
- Explain how to play the game and what the participants would experience (e.g. hearing periods of loud noise)
- Instruct participants to verbalize their actions/thoughts while playing, remind them of rules
- Answer questions before and during the games
- Adjust the game as necessary using timers, requesting status reports, issuing radio failures, otherwise modifying rules to elicit required speech from participants
- Compensate participants, schedule next session

Technical Assistant responsibilities included:

- Check and prepare headsets, recording software
- Test noise level at the beginning of each recording day
- Monitor recording throughout the game
- Save recording files and check metadata after each game
- Back up and upload recordings at the end of each day

Participant and session management were facilitated by the use of a custom web interface that allowed game masters to look up enrolled participants in the database by name, PIN or email address. Game masters then used the interface to assign participants to a given session, generate required session metadata, and assign the appropriate background noise recordings for that session. The interface was designed for ease of use by non-technical game masters to reduce the likelihood of data entry errors. During the recording session, game masters could also use the interface to log timestamped notes about any unusual or noteworthy occurrences that took place during the recording session.

While the game master was setting up the recording session and preparing participants for their game, the technical recording assistant prepared and checked the recording equipment and loaded the background audio recording designated for the first game of the session.

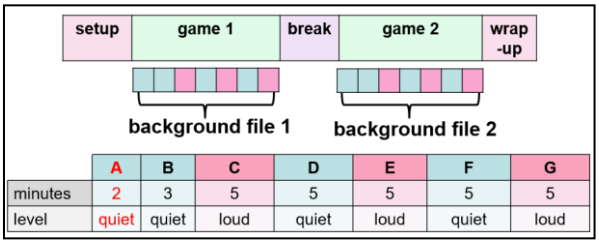
After the recording session ended, the technical assistant verified that the recordings had been saved to the local computer and that metadata was complete and accurate. Recordings were then uploaded to LDC's fileserver and also backed up to an external hard drive.

Participants were compensated in person at the end of each recording session using prepaid, reloadable debit cards. Immediate compensation after each session helped to encourage repeat participation and allowed for efficient payment tracking. To maximize the number of speakers with at least 2 hours of speech in the collection, we also offered bonus compensation after successful completion of the second recording session. Participants were also encouraged to sign up for their next recording session immediately following the end of the current session, which helped with participant retention.

4.3.2 Session Design

Each recording session included two, 30-minute games and each game used a background file that was unique for that pair of participants. Each 30-minute recording alternated between 5-minute quiet and loud intervals in which the background noise file is played at a quiet (0-14db) or loud (70-85db) volume. This alternating quiet/loud approach was found through testing to maximize the participants' range of vocal effort. Each 5-minute section was assigned letters for ease of reference.

The first 5-minute section is designated "AB" and begins with the participants' headphones off (no background noise) for 2 minutes (A) to get a baseline recording of the participant's typical voice. The participant then puts on their headphones and they stay on for the rest of the recording session. At the 2-minute mark quiet background noise begins (B), followed by 5 minutes of loud background noise (C), then 5 minutes of quiet (D), and so on. Babble noise, which consists of indistinguishable speech from multiple voices, was added to two of the three loud sections in addition to the operational background noise in order to elicit higher vocal effort from the speakers. Figure 1 illustrates the design of the recording sessions including the background noise conditions throughout each session.



	A	B	C	D	E	F	G
minutes	2	3	5	5	5	5	5
level	quiet	quiet	loud	quiet	loud	quiet	loud

Figure 1: Recording Session Design

4.3.3 Monitoring Noise Levels

While the objective of the SAFE-T Corpus collection was to generate recordings with speech exhibiting high vocal effort and the Lombard effect, it was necessary to ensure that participants were not subjected to unsafe levels of noise when trying to communicate with each other against the loud background noise piped into their headsets.

As noted above, loud background noise recordings were targeted to be 70-85 db. At 85 dB(A), speakers would need to shout to be heard by someone an arm's length away, making communication difficult. SAFE-T recording sessions included 1 hour of active game playing, during which participants heard loud background noise for no more than 5 minutes at a time for a total of up to 30 minutes per recording session. This level of exposure to loud noise is well within the Occupational Safety and Health Administration permissible exposure limit of 8 hours of noise at 90 dB(A) (OSHA, 1970).

The noise levels were monitored each day by using the generate white noise function in Audacity. Pre-set sound levels were measured by placing an Extech 407750 sound level meter with a mic attached within the headsets, to ensure that noise would be no greater than 90 dB(A) at its peak level.

5. Collection Infrastructure

5.1 Collection Platform

The collection platform consisted of a workstation, a digital audio interface, an analog matrix mixer, a backup drive, and four headsets. The platform was designed with several principles in mind:

- To allow two game participants to hear one another's speech, their own speech, and a background signal;
- To allow the system technician to hear the participants;
- To allow the game master to hear and speak to the game participants; and
- To capture the speech signal of each participant.

The collection room was a standard, rectangular multi-person office with additional carpeting installed to provide some amount of sound deadening. Foam acoustic panels were also attached to one of the walls to improve sound isolation from the adjacent offices. The recording equipment was placed on a single desk which was adjacent to the desk used for game play. Balanced, wired connections were used between the headsets and the analog matrix mixer, and between the mixer and the digital audio interface.

5.1.1 Recording Platform Components

The recording platform included the following components.

- BeyerDynamic DT290 headsets: dynamic, hypercardioid microphone, closed back earphones. These headsets included a directional boom microphone to cut down on external noise, had a closed back design to seal out all room noise, and were economically priced. This design helped to ensure that subjects heard the audio directly from the mixer rather than from other paths.
- Lectrosonics DM1612 Analog Matrix Mixer: 16 inputs, 12 outputs. This mixer connected the headphones, microphones, and computer audio input/output devices and allowed for both mixing and routing of audio between multiple sources and multiple targets. This enabled signal levels to be adjusted for all inputs and outputs, as well as

separation between signals sent to the headsets versus signals written to disk. The matrix mixer included microphone preamplification, crosspoint amplification, and output amplification and attenuation.

- Digigram VX882e PCI-E Audio Interface: 8 line level analog channels I/O, 8 AES/EBU digital channels I/O, Low Latency ASIO driver. The interface provided simultaneous signal capture and playback, acting as a bridge between the matrix mixer and the computer. The audio interface handled multichannel audio with low signal latency, which was important because the participants need to be able to hear one another without any delays that could interrupt natural speech behavior. We chose the Digigram VX882e because it is a very stable design with a long track record of successful audio processing and recording. The VX882e has balanced inputs and outputs, high quality filters, a very stable clock, and the ability to capture audio from multiple channels without dropping samples.
- HP Z6 Windows 10 workstation: 8-core Xeon CPU, 32GB RAM, 512GB SSD, 4TB External RAID. This workstation ran the audio capture/playback software, the matrix mixer configuration software, and the audio interface drivers.
- The Audacity 2.3.2 software package was used to handle background noise playback and clean channel capture. This software provided a straightforward user interface, good compatibility with the operating system and device drivers, and the ability to capture and playback audio simultaneously.
- The SoX v14.3.1 software utility was used to handle audio post processing. This software allowed for the batch processing of the audio files and was used to mix the background files with the clean channel files.
- The Lectrosonics Matrix Mixer API and Control Panel were used to handle signal routing and gain manipulation.

Figure 2 shows the overall design of the collection platform.

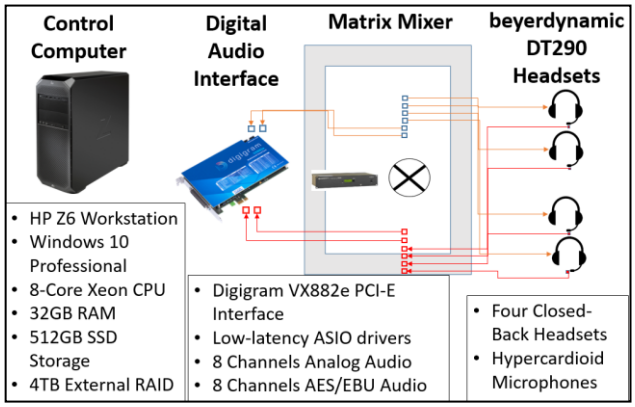


Figure 2: Collection Platform Design.

5.2 Noise Types, Background Types and Babble

5.2.1 Background Noise Files

Background audio files containing a variety of noise conditions were played into the participants' headsets during the 30-minute game sessions. Each background file was comprised of multiple 5-minute sections, each containing either silence, quiet background noise (0-14dB), loud background noise (70-85dB), or loud background noise with added speech babble (also 70-85dB). Table 1 summarizes the specifications for each portion of the background file recordings.

Section	Noise Type
AB	Silence and Quiet
C	Loud with Babble
D	Quiet
E	Loud without Babble
F	Quiet
G	Loud with Babble

Table 1: Background File Sections

5.2.2 Background Noise Collection and Auditing

The background files were designed to elicit variable vocal effort from participants and to reflect operational noise conditions. To achieve this, we collected real world noise samples from the web. Over 1500 recordings were manually scouted and collected, roughly equally divided into three noise types: event noise (e.g. sirens), vehicular noise (e.g. car motors), and environmental noise (e.g. HVAC systems).

After reviewing the collected background recordings with NIST, it was decided that all background files should use event noise, since this type was the most representative of first responder situations; most of the collected event noise samples were from amateur recordings of real world emergency scenes. It was also important that the background noise recordings did not contain discernible speech, to avoid complicating the SAD and ASR evaluation tasks with speech from non-target speakers. Therefore, the event type background noise recordings were manually audited for presence of discernible speech, and speech segments were excluded.

5.2.3 Background File Creation and Assignment

Background files were generated by building component files from the collected event noise recordings and then concatenating those individual audio files into a single background file as follows. A script randomly selected background event noise recordings from the tracking database until it reached a total duration of 5 minutes. The selected recordings were then processed with SoX and concatenated into a single 5-minute section, designated as either quiet or loud depending on its position in the background file as a whole. (The exception is the first 5-minute section, AB, which consists of 2 minutes of silence followed by 3 minutes of quiet noise; this is used to establish a baseline recording of the participant's voice without any background noise.) The gain for each section was then normalized to establish the loud or quiet noise condition as required, at -27dBFS for the quiet condition and -3dBFS for the loud condition. A second script then created the full background audio files by concatenating the

sections in their designated order (AB, C, D and so on). Information about each component file, section and complete background audio file were stored in separate tables in a centralized tracking database. The tracking database tables made the relationships between source audio, component files, segments, and background files explicit, so that it would be possible to recreate any of the component or background audio files if necessary.

Background files were assigned to participants in each session during session setup, as described in Section 4.3.1. Both of the speakers in a session were assigned the same background recordings, and files did not repeat for a given speaker across different recording sessions. The same background recordings could repeat across different speakers.

5.3 Signal Chain and Noise Inputs

The matrix mixer was used to route the signal between participant microphones, the digital audio interface and participant headsets. The matrix mixer had 3 stages:

1. Input stage, which includes up to 60 dB of preamplification;
2. Matrix stage, which allowed the signal to be routed, mixed, and have gain amplified/attenuated between 16 inputs and 12 outputs;
3. Output stage, which allowed for gain amplification and attenuation.

The Digigram audio interface had 8 line level analog inputs and 8 line level analog outputs. The background files were played through the Digigram audio interface output with 0dB gain/attenuation. The Digigram audio interface was connected to matrix mixer inputs set to 0dB gain.

The inputs were routed to both speakers' headsets with +3dB gain; the routing included both the matrix crosspoints for the two headsets and the amplifier stage of the mixer connected to each headset. The amplifier stage of the mixer was set to 10dB gain for each headset. This was done in order to set up multiple paths between the microphones and input signals (background files) and the headphones and capture files. We wanted to be able to route the signals so that we would have clean channel recordings from the microphones while simultaneously providing a noisy signal output to the participant headphones.

To elicit variable vocal effort, participants heard a combination of sounds through their headset while playing each game. Through the headset, the participant heard the designated background noise recording, their game partner's (and occasionally the game master's) voice, and their own voice (sidetone). After extensive testing at LDC and review by NIST, we established two noise levels and signal mixes for the headset inputs to produce the desired vocal effort. For the quiet sections, the output of the headphones ranged from 0 dB(A) to 14 dB(A), and consisted of 68% background noise, 16% partner's speech, and 16% sidetone. For the loud sections, the output of the headphones ranged from 70dB(A) to 85dB(A), and consisted of 86% background noise, 7% partner's speech, and 7% sidetone. The sidetone was captured through the participant's microphone and attenuated at the matrix mixer. In all cases, the mixture of background noise, game partner's speech and sidetone was done at the matrix mixer. Effective levels were dependent on actual signal activity at that moment, i.e., if no one was speaking, the signal heard through the headset would consist of 100% background noise.

5.4 Data Flow

The background file was played through two matrix mixer input channels, which were mixed with the microphone input and routed to the earphones. The microphone output, background noise file and babble file then went into the matrix mixer and were piped into the participants' headsets. Only the clean speech from each participant's mic was recorded, which was then stored as a recording file. Metadata for the session, the game and the participants was also written to the central database. Session metadata included timestamp, speaker ID, game number and background file ID. The clean single-channel speech recorded from a participant's microphone was used to manually produce reference transcripts, while the single-channel speech and background noise files were combined to create a mixed file that was used for OpenSAT training, development and test data.

6. Transcription

A portion of the collected data was earmarked for transcription. Twenty-five hours were selected as test data: 5 hours each for development and evaluation data for OpenSAT 2019, and an additional 15 hours to be used as test data in future OpenSAT evaluations. Another 97 hours of audio was selected for training data transcription. Development and evaluation transcript selections consist of four or five 3-minute snippets selected from the six, five-minute sections of each 30-minute single-channel recording. Training data transcripts consist of full 30-minute single-channel recordings.

All audio files used by transcribers were the single-channel participant recordings captured via the subject's close-talking microphone. Automatic speech activity detection (Ryant, 2013) was used to segment the audio prior to verbatim transcription of the primary speaker's speech. Speech from the game partner picked up by the primary speaker's head-mounted microphone was treated as background speech and was labeled as such but not transcribed. All transcripts were subject to multiple manual quality review passes and corpus-wide sanity checks, as described below.

6.1 Test Data Transcription and Quality Review

Test data (both development and evaluation) was transcribed to a Careful Transcription (CTR) standard (Glenn et al., 2010). Automatic SAD was used to create initial speech segments; transcribers then manually corrected automatic segmentation, adding or removing segments as needed and adjusting segment boundaries. Background speech was separately segmented (i.e. diarized) and background noise was also segmented. Transcribers produced a careful verbatim orthographic transcript for each speech segment, including indication of speaker noises like breath and cough, filled pauses, partial words, speaker restarts and other disfluencies. Markup was added for acronyms, proper nouns, spoken letters, foreign words, mispronounced words and other common phenomena including difficult-to-understand regions. A

complete second review pass was conducted by senior transcribers to verify the accuracy and completeness of the recording in its entirety, including both the transcript and its segmentation.

6.2 Training Data Transcription and Quality Review

Training data utilized a Quick Transcription (QTR) standard. QTR was designed to efficiently produce a verbatim transcript with minimal markup; this standard was selected for training data in order to increase the amount of transcribed speech available given a fixed timeline and budget. QTR segments were defined via automatic SAD without any manual correction. Background noise and background speech were not separately segmented; instead, primary speaker segments containing discernable background speech had a <*background*> tag inserted in the transcript itself. A special tag <*extreme background*> was used for background speech that was loud enough to compete with the primary speaker's voice. Unlike CTR, the QTR transcripts did not include special treatment of speaker noises and included limited markup for various speech and orthographic phenomena. QTR second passing involved listening to each segment in isolation and checking that its transcription was complete and accurate.

Prior to delivery, all transcripts, both CTR and QTR, were automatically checked for badly formatted tags, illegal characters, digits not spelled out, spacing issues and other common markup errors.

7. Preparing Data for Use in OpenSAT

All audio was delivered as single-channel, 48KHz 16-bit mono flac files. OpenSAT 2019 required that we produce mixed files consisting of single-channel speech recordings mixed with background noise recordings at a reduced level. Although babble noise was heard by participants during recording sessions (since it proved very effective in producing the desired vocal effort), it was excluded from the mixed files used in OpenSAT since babble noise does not represent the kind of noise generally present during emergency situations. The mixed files were created by matching the clean channel recordings with their corresponding background files and mixing them using the SoX gain function.

All transcripts were released in a simple tab-delimited format with UTF-8 encoding. In addition to transcripts and audio, release packages include speaker and session metadata. Audio recording and transcript file names reference their associated metadata, following this convention:

<PIN>_<YYYYMMDD>_<hhmmss>_part<1|2>_
<AB|C|D|E|F|G>_<partition>

7.1 Reduced Background Noise Level in Mixed Files

A goal of the corpus was to create speech recordings to mimic public safety communications including realistic background noise that would be reasonably challenging for system developers in the OpenSAT 2019 evaluation. Through testing, NIST found that mixing the speech recordings with the background files at the full level heard by the speakers themselves made the loud sections too challenging for use as evaluation data. It was therefore decided that the background noise recordings should be mixed with the speech recordings at a reduced level. After extensive testing, the background file recordings were reduced such that the loud sections had a peak level of -12dBFS RMS.

To prepare the mixed files for use in OpenSAT, the background file signal levels were reduced in the loud sections to better match the signal levels of the clean channel recordings. We used SoX to reduce the levels, then combined the background files with clean channel recordings using the "sox --combine mix-power" command.

The original background files consisted of silence followed by alternating quiet (-27dBFS) and loud (-3dBFS max) sections; this is the version of the background file that was played through the participant headsets during the recording session. The reduced level background files consisted of silence followed by alternating quiet (-36dBFS max RMS) and loud (-12dBFS max RMS) sections, which were normalized, i.e. the amplitude of the digital audio was scaled down relative to the max RMS level.

8. Challenges and Solutions

8.1 Crosstalk Speech

It was a known risk given the collection protocol that the game partner's (and occasionally the game master's) speech would be audible on the primary speaker's mic and thus present in the single-channel speaker recordings. This risk was mitigated through game manipulation (e.g. addition of the physical barrier, instructing subjects to take turns speaking). The microphones used for the collection were also selected to minimize capture of crosstalk to the extent possible. The transcription methodology was also designed to manage this risk by flagging background speech in all training data, and by segmenting and diarizing background speech in the development and evaluation data.

The decibel level of the crosstalk picked up on the primary speaker's mic and present in the recording was not measured, but is at a noticeably lower level than that of the primary speaker and is easily distinguishable from the primary speech.

8.2 Eliciting High Vocal Effort

A major challenge when designing the collection protocol for SAFE-T was to balance the goal of eliciting high vocal effort, Lombard effect and urgent speech with the comfort of study participants including their ability to play the game effectively under challenging recording conditions. It was important for participants to really engage in the game so that their speech would mimic the properties of operational speech as closely as possible given a simulated setting. We also needed participants to return for multiple sessions so that we would have sufficient speech from each individual to meet our targets. As such, we put effort into making the game as enjoyable and engaging as possible. We created a study competition to encourage urgent speech by posting a leader board of the total number of victims rescued by each team. We also extensively tested background noise levels and durations with the participant experience in mind; for instance, we found that using 5-minute intervals of loud background noise not only elicited a better range of vocal

effort, but also made the participants more comfortable than longer intervals. The vast majority of study participants (87%) completed the minimum of two required recording sessions, with 38% completing the maximum of four. Many participants provided positive feedback about their experience and some even voluntarily spread the word about the project to other potential participants. All of these factors made it possible to not only meet but to exceed our collection goals within the time and budget constraints of the collection.

9. Conclusions

The SAFE-T Corpus is a new resource for speech analytic research in the Public Safety Communications domain, comprising over 300 hours of speech from over 100 speakers. A portion of the data has been manually transcribed. The amount of speech vs. non-speech, the degree of interaction between speakers, and the level of vocal intensity present in the collected audio satisfy collection requirements and reflect key properties of operational data in the public safety domain. The final makeup of the SAFE-T Corpus collection is summarized in Table 2 below.

	Goal	Completed
Collected Audio Hours	291	330
Unique Speakers	100	115

	Audio (Hours)	Transcript (Hours)
Training data released to date	131	50
Evaluation data released to date	5	5
Development data released to date	5	5
Training data available for future evals	174	47
Test data available for future evals	15	15

Table 2: SAFE-T Corpus Totals

The initial set of collected and transcribed data has been released to performers in the OpenSAT 2019 evaluation, and additional recordings and transcripts will be utilized in future OpenSAT evaluations. The full SAFE-T corpus will be published in LDC's catalog after the data is no longer sequestered for use in evaluations, along with the OpenSAT training, development and test data sets.

10. Acknowledgements

The authors gratefully acknowledge the contributions of Frederick Byers and the team at NIST for their guidance and feedback on the SAFE-T Corpus design.

11. References

11.1 Bibliographical References

Bagwell, C., Sykes, R., Giard, P. (2010) SoX 14.3.1. http://sox.sourceforge.net/sox.html https://sourceforge.net/p/sox/code/ci/master/tree/AUTHORS

Baur, C., Rayner, E., Tsourakis, N. (2014) Using a Serious Game to Collect a Child Learner Speech Corpus. In: Ninth International Conference on Language Resources and Evaluation. In Proceedings of LREC 2014, Reykjavik Iceland

Glenn, M.L., Strassel, S.M., Lee, H., Maeda, K., Zakhary, R., Li, X. (2010) Transcription Methods for Consistency, Volume and Efficiency In Proceedings of LREC 2010 Valletta, Malta

Gravano, A., Hirschberg, J. (2011) "Turn-taking cues in task-oriented dialogue," Comp. Speech and Language, 25(3), pp.601-634, 2011. http://www.cs.columbia.edu/speech/games-corpus/

Lanzing, K. Flash Point Fire Rescue. 999 Games, Hobby Japan, Indie Boards and Cards, MINDOK, 2011. Board Game.

NIST Open Speech Analytic Technologies 2019 Evaluation Plan. 2019. Accessed November 30, 2019. https://www.nist.gov/document/opensat19evaluationplanv43-28-19pdf

Occupational Safety and Health Administration. (1970). The Occupational Safety and Health Administration's (OSHA's) Noise standard (29 CFR 1910.95). Retrieved from https://www.osha.gov/Publications/laboratory/OSHAfactsheet-laboratory-safety-noise.pdf

Ryant, N. (2013). "LDC HMM Speech Activity Detector (v.1.0.5)." LDC, University of Pennsylvania.

11.2 Language Resource References

Schmidt-Nielsen, A., et al. (2000) Speech in Noisy Environments (SPINE) Training Audio LDC2000S87. Web Download. Philadelphia: Linguistic Data Consortium