Voices Obscured in Complex Environmental Settings (VOiCES) corpus

Colleen Richey[2]* and *Maria A.Barrios*[1]*, *Zeb Armstrong*[2], *Chris Bartels*[2], *Horacio Franco*[2], *Martin Graciarena*[2], *Aaron Lawson*[2], *Mahesh Kumar Nandwana*[2], *Allen Stauffer*[2], *Julien van Hout*[2], *Paul Gamble*[1], *Jeff Hetherly*[1], *Cory Stephenson*[1], and *Karl Ni*[1]*

[1]Lab41, In-Q-Tel Laboratories, Menlo Park, CA 94025
[2]SRI International, Menlo Park, CA 94025
* Equal author contribution

`colleen@speech.sri.com, mbarrios@iqt.org`

Abstract

This paper introduces the Voices Obscured in Complex Environmental Settings (VOiCES) corpus, a freely available dataset under Creative Commons BY 4.0. This dataset will promote speech and signal processing research of speech recorded by far-field microphones in noisy room conditions. Publicly available speech corpora are mostly composed of isolated speech at close-range microphony. A typical approach to better represent realistic scenarios, is to convolve clean speech with noise and simulated room response for model training. Despite these efforts, model performance degrades when tested against uncurated speech in natural conditions. For this corpus, audio was recorded in furnished rooms with background noise played in conjunction with foreground speech selected from the LibriSpeech corpus. Multiple sessions were recorded in each room to accommodate for all foreground speech-background noise combinations. Audio was recorded using twelve microphones placed throughout the room, resulting in 120 hours of audio per microphone. This work is a multi-organizational effort led by SRI International and Lab41 with the intent to push forward state-of-the-art distant microphone approaches in signal processing and speech recognition.

Index Terms: corpus, speech recognition, speaker recognition, data collection, LibriSpeech

1. Introduction

SRI International and Lab41, In-Q-Tel, are proud to release the VOiCES Obscured in Complex Environmental Settings (VOiCES) corpus, a collaborative effort that brings speech data in acoustically challenging reverberant environments to the researcher. Clean speech was recorded in rooms of different sizes, each having distinct room acoustic profiles, with background noise played concurrently. The corpus contains the source audio, the retransmitted audio, orthographic transcriptions, and speaker labels. The ultimate goal of this corpus is to advance acoustic research by providing access to complex acoustic data. The corpus will be released as open source, Creative Commons BY 4.0, free for commercial, academic, and government use.

Datasets for speech research are typically expensive, limited in scope, and behind paywalls. Synthetic data can be created by superimposing audio samples from datasets of isolated speech and noise and using software to generate reverberation[1]. Unfortunately, these techniques do not accurately represent the acoustics of real-world environments and dynamic noise. On the other hand, publicly available datasets collected in real environments often use few speakers[2]. Data competitions like CHiME have provided increasingly more re-

alistic data, though with limited number of speakers. Early CHiME datsets[3] were constructed by convolving clean speech with a *simulated* room response (based on measured data for 2 rooms), assuming the speaker to be 2m from the recorded audio. This signal was then mixed with the recorded multi-source background noise recorded in the rooms. Extended work[4] later added *simulated* location changes within a 20cmX20cm area and small 5 cm head movement translation. Integrated recording in real environments was introduced in later challenges[5, 6], but this included only 4 speakers in four different settings, recorded at close range microphony via 1, 2, or 6 microphones. Data for this year's CHiME challenge includes 40 speakers recorded at homes, using binaural microphones and microphone arrays placed in each room. In contrast, VOiCES includes 300 speakers, a range of distractor noise types, various types of microphes at a distance, and a rotation range of $180°$ for the foreground loudspeaker position. This article reports results from recordings done in two rooms. The full corpus will include additional rooms; these recordings are ongoing.

Successfully deploying speech and acoustic signal processing algorithms in the field hinges on access to realistic data. To this end, audio for the VOiCES corpus was recorded under conditions that better represent real-use situations. These recordings provide noisy, reverberant audio with the intended purpose of promoting acoustic research including speech processing (speaker identification and acoustic detection, speech recognition), audio classification (event and background classification, speech/non-speech), and acoustic signal processing (source separation and localization, noise reduction, general enhancement, acoustic quality metrics). In the remainder of this paper, a detailed description of the VOiCES corpus is provided, including model baselines for automatic speech recognition and speaker identification. Section 2 describes the collection effort itself, Section 3 provides some insight into the statistics of the dataset, and Section 4 outlines model baselines that were run on the dataset. The corpus will be available on Amazon Web Services, where details on use cases and a download link will be provided.

2. Dataset Collection

The main focus when developing the VOiCES corpus was to provide an open-source dataset centered on distant microphone collection under realistic conditions. Pre-recorded foreground speech and background noise were played in two furnished rooms with different acoustic profiles (reverberation, HVAC background, echo, etc.) and was recorded by 12 distant microphones. Recording rooms were windowed and carpeted, with mostly bare walls and a bare ceiling, furnished with tables and
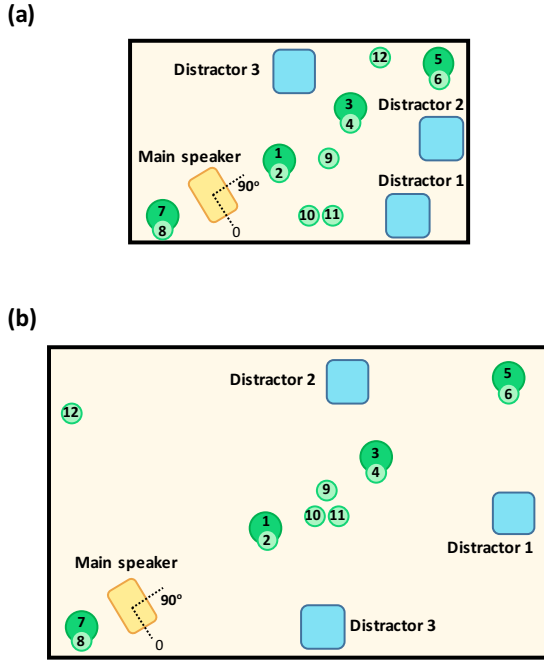
Figure 1: *Microphone and loudspeaker configuration (not to scale) used for recording sessions in (a) room 1 (146" x 107") and (b) room 2 (225" x 158"). The foreground loudspeaker (at its $90°$ position), orange rectangle, was placed in a corner of the room, and speakers playing noise, blue squares, were placed with their cones directed toward the center of the room. Studio and lavalier microphones are shown as large (dark) and small (light) green circles; microphone ID and distance from foreground loudspeaker are listed in Table 1.*

chairs. Four recording sessions were held in each room: one for each distractor noise type (television, radio, or babble) played concurrently with the foreground speech, and one session with foreground speech only. One hour of only distractor noise or ambient room background noise was recorded at the end of each session. This resulted in over 120 hours of recorded speech per microphone, for a total of 374,688 audio files and 1440 hrs of recorded speech.

2.1. Audio Sources

The audio for foreground speech and distractor noise was selected from sources either in the public domain or under a creative commons attribution license that permits data derivatives and commercial use.

2.1.1. Foreground Speech

A total of 15 hours (3,903 audio files) were selected from LibriSpeech[7], a corpus of audiobooks in the public domain. All audio contains English read speech. Audio was taken from 300 speakers in the "clean" data subsets, with an even split between females and males. At least three minutes of speech were selected from each speaker, with at least one minute from three different book chapters - an amount sufficient for speaker identification tasks. LibriSpeech files use a sample rate of 16kHz, 16-bit precision, and Free Lossless Audio Codec (FLAC) encoding. Selected files were corrected for DC offset, normalized

based on their peak amplitude, and converted to WAV format. The selected audio files were concatenated together with 2 seconds of intervening silence into a continuous audio file. The loudspeaker playing the foreground speech was on a motorized rotating platform. The order of the individual audio files was randomized, to guarantee that there was no correlation between a particular human speaker and a position of the loudspeaker. Signals to evaluate the room response were added at the beginning of each session. These included a steady tone, a rising tone, and a transient noise. The final concatenated source file was 19 hours long.

2.1.2. Distractor Noise

Audio was recorded under four different noise conditions: one without any added noise (ambient room noise only) and three with a distractor noise played simultaneously with the foreground speech. The distractor noises were television, music, or overlapping speech from multiple speakers (referred to here as babble). During recording sessions, the audio for television or music was played from a single loudspeaker; babble was played from three noise-dedicated loudspeakers. An extra hour of just distractor noise was recorded at the end of each session.

Television noise was selected from movies and television shows in the public domain[8, 9]. Audio from 76 videos was extracted in M4A format and converted to WAV with a 16kHz sample rate and 16-bit precision. Five-minute excerpts were chosen from each audio file and each excerpt was normalized to its peak amplitude. Depending on the length of the source audio, 5 to 8 excerpts were taken from each movie or show, randomized, and concatenated into a single 20-hour audio file.

Music noise was selected from the MUSAN corpus[10]. All music files are in the public domain or under a Creative Commons license. Any music files having no derivative (ND) or non-commercial (NC) license restrictions were omitted from the sample set. The music files were randomized and concatenated into a single 20-hour audio file. Due to the large variability in signal amplitudes for different genres of music, the concatenated audio file was run through the compander tool in the SoX audio utility, combining compression and expansion of the signal dynamic range. This ensured a more uniform music volume throughout the recording sessions and a more consistent signal-to-noise ratio.

Babble noise was constructed using the "us-gov" subset of the MUSAN corpus[10]. This subset contains audio recording excerpts of various US government meetings; all are in the public domain. Each excerpt is about 5 minutes long and was normalized to its peak amplitude. Babble tracks were constructed by randomizing and concatenating together meeting excerpts into 20-hour audio files and then mixing three audio files into one. Three babble tracks were created and were played out of three noise-dedicated loudspeakers (i.e. at least nine overlapping speakers) simultaneously with the foreground speech.

2.2. Recording Setup

Two different rooms were used for recording: room-1 with dimensions 146" x 107" (x 107" height) and room-2 with dimensions 225" x 158" (x 109" height). Twelve microphones were placed in strategic locations throughout the room: 7 cardioid dynamic studio microphones (SHURE SM58), 4 omnidirectional condenser lavalier microphones (AKG 417L), and 1 omnidirectional dynamic lavalier microphone (SHURE SM11). Paired studio and lavalier microphones were placed at four different positions: (1) Behind the foreground loudspeaker, (2) on a table

Table 1: *Microphone type, location, distance from foreground loudspeaker (s) and height (h) for room-1 and -2 configurations.*

Mic ID (type)	Location	Room-1 (s, h)	Room-2 (s, h)
01 (studio), 02 (lavalier)	near on table	(38", 42")	(80", 39")
03 (studio), 04 (lavalier)	far on table	(72", 42")	(131", 39")
05 (studio), 06 (lavalier)	across room	(119", 70")	(228", 70")
07 (studio), 08 (lavalier)	behind loudspeaker	(29", 70")	(29", 70")
09 (lavalier)	partially obstructed, table	(58", 28")	(109", 25")
10 (lavalier)	on ceiling, clear	(75", 105")	(128", 105")
11 (lavalier)	on ceiling, fully obstructed	(75", 106")	(128", 106")
12 (lavalier)	fully obstructed, wall	(130", 12")	(116", 10")

directly in front of the foreground loudspeaker, (3) on a table in front of the foreground loudspeaker at a farther distance than (2), and (4) across the room from the foreground loudspeaker. The remaining four lavalier microphones were placed in other locations in the room, fully or partially obstructed by a physical barrier. Distances between the foreground loudspeaker and microphones are listed in Table 1. All audio was played on high-quality speakers; one speaker was reserved for foreground speech, and three others were used to play distractor noise. A schematic of speaker and microphone placement in both rooms in shown in Figure 1.

The foreground speaker was placed 43" from the floor on a robotic platform that automatically rotated the position of the foreground speaker by ten degrees every hour, spanning a total of 180 degrees. The rotating platform's step motor was sufficiently shielded to prevent recording background noise from the motor movement. The motivation to have a non-static audio source was to emulate common human behavior that occurs during conversations such as head movement or walking, that is not captured in other datasets.

A PreSonus StudioLive RML32AI digital mixer and PreSonus Capture recording software were used to play and record the audio. A sound pressure meter, placed close to microphone 01, was used to measure the playback audio and adjust volume levels on the PreSonus mixer for both the foreground audio (~65 dB) and distractor noise (~50 dB). All channels were sample synchronous. Each recording session lasted 20 hours (19 hours of foreground speech and 1 hour of only distractor or ambient noise). The recording sessions were segmented according to the source files from LibriSpeech, yielding 1440 hours of audio (347,688 audio files) across all microphones and sessions. Audio was recorded with a 48kHz sample rate and 24-bit precision in WAV format with PCM encoding, and is also available in 16kHz and 16-bit precision in WAV format. The corpus also contains the source audio files (16kHZ sample rate, 16-bit precision, WAV format).

3. Data Statistics

To obtain an assessment of the statistics of the corpus, the duration, minimum and maximum amplitude, root mean square (RMS) energy, and signal-to-noise ratio (SNR) were calculated for all audio files in the corpus. Statistics were calculated using a combination of the SoX utility and SRI's in-house utilities.

The average and median duration for all data subsets is 15.62s and 15.97s, respectively, with a standard deviation of 1.91s. This is evidence that the automatic audio segmentation worked correctly and that we can directly compare noisy files with source files.

The RMS, measuring the amplitude of the audio file relative to the digital system's maximum level (with maximum value at

0 decibels relative to full scale - dBFS), was consistent across the various subsets. Average values were measured between -27.0 and -27.5 dBFS, indicating the playback volume was consistently set for all recordings.

The minimum and maximum amplitudes represent the lowest and highest amplitude for samples in a given audio file, on a normalized scale of ± 1. These ranged between -0.5 to 0.5 across all data subsets, showing reasonable use of the digital recording systems levels. The average minimum and maximum amplitude levels for the source audio were -0.93 and 0.91.

The signal-to-noise ratio (SNR) measures the strength of a primary signal relative to the background noise. Differences in SNR were evident between rooms and distractor noises, and degraded with increasing distance between the foreground loudspeaker and microphone. The average SNR for audio recorded in room-1 and -2 was 22.19 dB and 19.50 dB, respectively.

Table 2 shows the calculated SNR for audio recorded under different noise conditions as compared to the source audio's SNR. The SNR significantly degrades for audio recorded at a distance in a real acoustic environment, even without distractor noise. A decrease of 18 dB was observed for this case. The addition of noise further decreases the SNR. The SNR for microphones close to and behind the foreground loudspeaker was 22.3 dB, and for those at mid- and far-distance, it was 20.5 dB.

Table 2: *Measured SNR for the source audio and audio recorded at a distance with and without distractor noise.*

	Source	No distractor	Music	TV	Babble
SNR	41.7	23.6	19.2	22.2	18.4

4. Model Baselines

SRI's in-house automatic speech recognition (ASR) and speaker identification (SID) systems were used to examine the recorded data. This provides data validation for analytics and a point of reference for future model implementations.

4.1. Automatic speech recognition (ASR)

The ASR system was run on a subset of data: audio from lavalier microphones when the foreground loudspeaker was positioned at $90°$ (directly aligned with microphones on table). The ASR system was built using the Kaldi Speech Recognition Toolkit [11]. It uses filterbank features and a time delay neural network (TDNN) and was trained on 500 hours of segmented English speech, which included data collected under DARPA's Translation Systems for Tactical Use (TRANSTAC) program and SRI proprietary data. Training audio is included twice, once in its original form and a second with artificially added rever-

beration. Because no full test or development subset of data from LibriSpeech is included in the VOiCES corpus, a direct comparison with published ASR results using LibriSpeech is not possible. It is possible, however, to make a rough comparison with results using the dev-clean LibriSpeech dataset. Published results for this subset achieved 4.9% and 7.8% word error rate (WER), for models trained on LibriSpeech and on the Wall Street Journal data, respectively[7]. The SRI system achieved a 9.3% WER.

Table 3 shows the WER when the foreground speaker is at $90°$ (centered) as a function of distractor noise. Results show a sharp increase in WER for data recorded in realistic acoustic environments. The WER for audio recorded by distance microphones with no added distractor noise is 19.0% - more than double the WER on the source audio. Added distractor noise degrade the performance further. The worst performance is on audio with babble noise, as this type of noise contains only speech and easily confuses the ASR system.

Table 3: *WER as a function of distractor noise type for room-1 and room-2 (mics 02, 04, 06, and 08), with foreground loudspeaker at $90°$, obtained from in-house SRI ASR system.*

	Source	No distractor	TV	Music	Babble
WER	9.3	19.0	29.3	27.6	33.0

In general the ASR performance is dependent on the distance between the foreground loudspeaker and microphone, and on individual room acoustics, as depicted in Figure 2. Results are shown for microphones 02, 04, 06 in both rooms when the foreground loudspeaker is at $90°$. There is an increase in WER with increased distance between the microphone and foreground loudspeaker. Differences in WER for microphones in room-1 and room-2 that are at comparable distances show the effect of each room's acoustic environment.
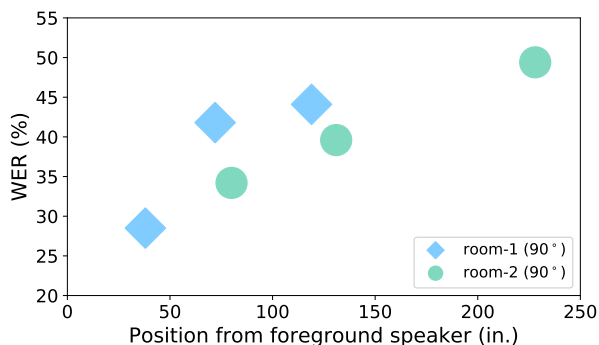

Figure 2: *The WER performance is affected by distance from the foreground loudspeaker, as well as room acoustic profile.*

4.2. Speaker identification (SID)

A state-of-the-art SID system from SRI was run on the VOiCES corpus[12]. The model used is a Universal Background Model (UBM) identity vector (i-vector) based system [13, 14], with a probabilistic linear discriminant analysis (PLDA) [15] as back-end classifier. A gender-independent PLDA was used to compute the scores of the speaker recognition system. The model was trained using the PRISM dataset [16]. The equal error rate (EER), describing the value where false positives equal false

negatives, is used as the metric for the SID system performance. For our experimental setup, we ensured enroll and test audio segments corresponded to different book chapters from the original corpus. Speech segments were on average 14s long for both enroll and test subsets. Results are shown for microphones 01 and 02 (Close), microphones 03 and 04 (Mid), and microphones 05 and 06 (Far).

Table 4 shows the impact of microphone distance on the SID performance. In this experiment, enrollment was performed on clean source data, and the EER is shown when testing on a variety of distant conditions. In order to highlight the effect of distance alone, no distractor noises were used. We observe that the EER of this UBM-IV system doubles when comparing the source audio (5.72%) to audio from the close microphones for both rooms (10.7%-10.9%), and it almost triples for the far room microphone (15.1%-16.6%).

Table 4: *Impact of microphone distance on the performance of the UBM-IV speaker recognition systems EER (%).*

Mics	Source	Close	Mid	Far
Rm1	5.72	10.7	13.0	15.1
Rm2	5.72	10.9	13.2	16.6

Table 5 shows the effect of distractor noise on SID performance. In order to mimic a realistic test case, speakers were enrolled using a recording from the close lavalier microphone (Close) in room-1 with no distractor noise. The test segments originate from all microphones and were recorded in room-2 with different types of background noise. We observe that distractor noise degrades the EER by 2% absolute for music and television and 3.5% absolute for babble. This is perhaps because it is very speech-like, but also possibly because babble was the only distractor played out of three separate loudspeakers.

Table 5: *Impact of distractor noise on the performance of the UBM-IV speaker recognition systems in terms of EER (%). Each condition has above $18k/2.8M$ target/impostor trials.*

Distractor	No distractor	TV	Music	Babble
UBM-IV	17.2	19.3	19.2	20.9

5. Conclusions and Future Work

The VOiCES corpus provides audio data that closely resemble acoustic conditions found in real recording environments - distant microphones, background noise, and reverberant room acoustics. The corpus can serve as a test and development set for research in the areas of speech and acoustics. It will enable the development of robust acoustic models that can better perform in the wild. By making the corpus publicly available, SRI International and Lab41 hope to promote and advance acoustic research on event and background detection, source separation, speech enhancement, source distance and sound localization, speech activity detection, as well as speaker and speech recognition. Data presented here correspond to phase I data collection. The corpus will be augmented with further data collection in phase II, that will include additional rooms and more challenging distractor noise profiles.

6. References

[1] T. H. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," in *2010 IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, 2010, pp. 90–100.

[2] Q.Jin, T.Schultz, and A.Waibel, "Far-field speaker recognition," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, 2007, pp. 2023–2032.

[3] J. Barker, E. Vincent, N. Ma, C. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, May 2013.

[4] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dec 2013, pp. 162–167.

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHIME speech separation and recognition challenge: Analysis and outcomes," *Computer Speech and Language*, vol. 46, pp. 605–626, Nov 2017.

[6] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech Language*, vol. 46, pp. 535 – 557, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230816301231

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5206–5210.

[8] publicdomainmovies.net. (2018) Public domain movies. [Online]. Available: http://publicdomainmovies.net/

[9] P. S. E. LLC. (2018) Public domain movies and TV shows. [Online]. Available: http://publicdomainmovies.org/

[10] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015. [Online]. Available: http://arxiv.org/abs/1510.08484

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[12] M. K. Nandwana, J. van Hout, M. McLaren, A. Stauffer, C. Richey, A. Lawson, and M. Graciarena, "Robust speaker recognition from distant speech under real reverberant environments using speaker embeddings," *Interspeech*, vol. Accepted, 2018.

[13] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[14] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Proc. Interspeech*, 01 2011, pp. 249–252.

[15] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*, Oct 2007, pp. 1–8.

[16] L. Ferrer, H. Bratt, L. Burget, H. Cernockyy, O. Glembeky, M. Graciarena, A. Lawson, Y. Lei, P. Matejkay, O. Plchoty, and N. Scheffer, "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," in *Proceedings of NIST 2011 workshop*, 2011.