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]Lab 41, In-Q-Tel Laboratories, Menlo Park, CA, USA, [2]SRI International, Menlo Park, CA, USA
* Equal author contribution

Abstract

We introduce the Voices Obscured in Complex Environmental Settings (VOiCES) corpus, a **freely available** dataset, recorded by far-field microphones in acoustically challenging reverberant environments. Clean speech was recorded in furnished rooms of different sizes, each having distinct room acoustic profiles, with background noise played concurrently. Audio was recorded using 12 microphones placed throughout the room, resulting in 120 hours of audio per microphone. These recordings provide audio data that better represent real-use scenarios. 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.

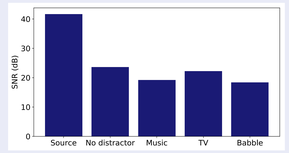
Dataset Details

Mic and loudspeaker placement in room-1 and sample photographs from both rooms. Circles represent mics (large – studio, small – lavalier), rectangles represent loudspeakers.

- ***Source Audio***: LibriSpeech[1] – 15 hrs (3,903 audio files), 300 speakers
  - 150 females, 150 males
  - 3-5 minutes of speech per speaker
- ***Distractor Noise***: (from MUSAN)
  - Television
  - Music
  - Babble (blend of 3 meetings)
- ***Background Noise***:
  - HVAC, fluorescent lights, room reverberation
  - external noise
- ***Simulated head movement***: motorized rotation platform with 180° span
- ***Multiple rooms***: carpeted and furnished, dimensions: 146”x107” & 225”x158”
- ***Various types of Microphones***:
  - 4 cardioid dynamic studio mics (SHURE SM58)
  - 7 omnidirectional condenser lapel mics (AKG 417L)
  - 1 omnidirectional dynamic lapel mic (SHURE SM11)
- ***Concurrent playback and recording***: PreSonus StudioLive RML32AI digital mixer
  - ~15dB difference between playback volume of foreground speech and noise, at mic 1 location
  - Recorded at 48kHz sample rate and 24-bit precision in WAV format
  - All channels are sample synchronous

Corpus Statistics

- 1440 hours of retransmitted distant audio
- 375K audio files with average duration of 15.6 s
- SNR degrades as a function of distance from foreground speaker, and varies between rooms



Measured SNR for the source audio and VOiCES data

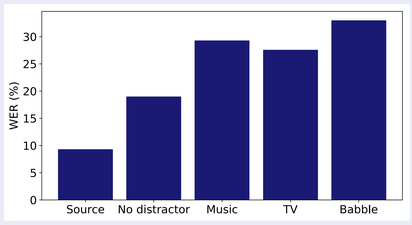
Corpus	Speakers	Hrs.	# mics	Noise	Room acoustics	Changing angle btw Speaker & mic	Mic type	Distant microphony	Public
VOiCES	300	1440	12 (20)	TV, music, babble	YES- real	YES	Studio, lavalier, MEMS	YES	YES
LibriSpeech	2484	1000	1	---	NO	NO	studio	NO	YES
WSJ			1	---	NO	NO	studio	NO	NO
CHiME-3*	12	~7	7	Café, street, public transport, pedestrian area	YES	NO	lavalier	NO	NO**
TED-LIUM	685	118	1	---	NO	NO	lavalier	NO	YES

* Description for real recordings on CHiME-3 dataset; dataset also includes simulated data, not tallied here
** CHiME-4, a subset of CHiME-3 is publically available

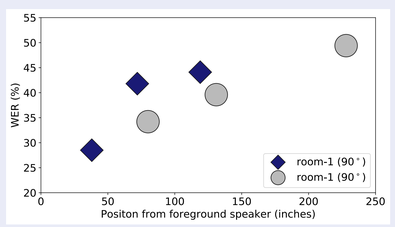
Model Baselines

SRI’s in-house automatic speech recognition (ASR) and speaker identification (SID) systems, and Lab41’s audio denoising were used to examine the recorded data. This provides data validation for analytics and a point of reference for future model implementations.

Automatic speech recognition (ASR)



Word error rate (WER) for varying noise type, with foreground speaker at 90°. Data from room-1 and -2 (mics 02, 04, 06, & 08)

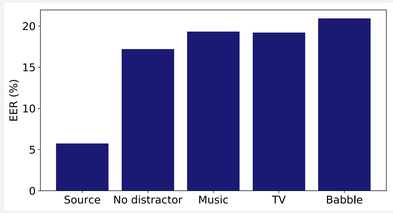


The WER performance is affected by room acoustics and the distance from the foreground loudspeaker.

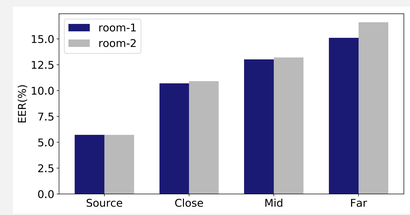
- Filterbank features and a time delay neural network (TDNN)
- Trained on 500 hours of segmented English speech (TRANSTAC + SRI proprietary)
- Training includes original audio with and without artificially added reverberation

Speaker identification (SID)

- UBM i-vector based system, with PLDA as backend classifier.
- Trained using PRISM dataset; Enroll and test segments from different sessions



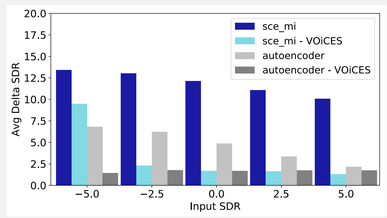
SID performance vs. noise type. Enrolled done using mic02 in room-1 (no distractor). All mics in room-2 are test segment.



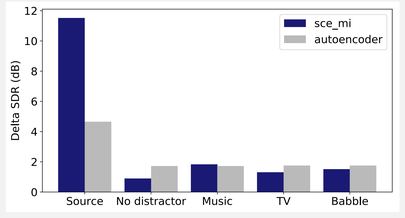
Impact of distance on SID EER(%) ; enrollment using source data; test segments in rooms with no distractor noise.

Audio denoising

- Two models: (1) Source-contrastive estimation with mask inference embedding (SCE_MI)
  (2) Convolutional denoising auto encoder
- Trained on synthetic mixtures of LibriSpeech and non-stationary noise from UrbanSound8k; mixes with SNRs ranging from -5 to 5 dB



Delta SDR as a function of input SDR for various models.



Enhancement in SDR by noise type.

How to Obtain the Corpus

- Additional information and instructions on how to download the corpus are available at : **voices.lab41.org**
- Includes force aligned transcripts from LibriSpeech
- Audio: 16kHz 16-bit WAV

Scan me

Conclusions and Future Work

- Large free speech corpus with realistic recording conditions
- Several controlled variables:
  - Room conditions, microphone distance, background noise
  - Angle between foreground loudspeaker and microphones
- Intended research areas include:
  - ASR, speech activity detection, Speaker ID, Speech enhancement
  - Source separation, event and background detection, source distance and sound localization
- Collection of two additional rooms with different acoustics to be released Oct 2018
  - Additional studio microphones to record lower SNR
  - Inclusion of MEMS microphones