The Voices Obscured in Complex Environmental Settings (VOiCES) Corpus

Formerly: The Speakers in the Room (SITR) Corpus

Presentation Number: 3pSC4

Aaron Lawson[1], Colleen Richey[1], Zeb Armstrong[1], Martin Graciarena[1], Karl Ni[2], Todd Stavish[2], Cory Stephenson[2], Jeff Hetherly[2], Paul Gamble[2], Maria Barrios[2]

[1]SRI International, Menlo Park, CA, USA [2]Lab41, Menlo Park, CA, USA

Introduction

- **Free** speech corpus
  - 12 far-field microphones at different **distances**
  - Different **room conditions** & Different **background noise**
- Retransmitted speech from LibriSpeech ASR Corpus
  - English read speech
  - 15 hours of source speech => 1440 hours of distant speech
- Freely available under CC 4.0 license
  - Download from Amazon Web Services

Corpus Collection

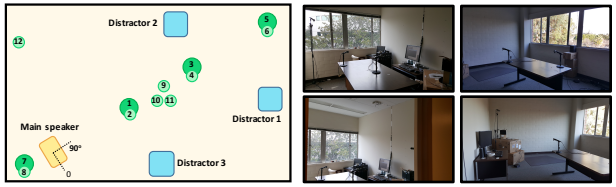
Source Audio

- Subset of LibriSpeech ASR Corpus
  - www.openslr.org/12
  - English audio books in the public domain
  - Selected 300 speakers (150 female, 150 male)
  - 3-5 minutes of speech per speaker
  - 4K audio files with average duration of 15.6 sec
- Selected 3 types of background noise
  - Music - from MUSAN: A Music, Speech, and Noise Corpus (www.openslr.org/17/)
  - Television - from movies and TV shows in the public domain
  - Babble speech – created from speech in MUSAN: A Music, Speech, and Noise Corpus

Recording Setup

- 2 rooms – small and large
  - 146”x107” & 225”x158”
  - Carpeted, furnished, windows, HVAC, reverberant
- Speech played from a single loudspeaker
  - On a platform that rotates 180 degrees

Figure 1: Schematic of microphone and loudspeaker placement in the larger room and sample photographs from both rooms. Larger circles are studio mics; smaller circles are lavalier mics. Mic 9 is partially obstructed on the table; mics 10 & 11 are attached to the ceiling, and mic 12 is on the wall, fully obstructed.



- 4 background noise conditions (“distractors”)
  - No added noise, music, TV, babble noise
- Music and TV noise played from single loudspeaker
- Babble noise played from 3 loudspeakers
- 12 microphones placed throughout each room
  - 4 cardioid dynamic studio mics (SHURE SM58)
  - 7 omnidirectional condenser lapel mics (AKG 417L)
  - 1 omnidirectional dynamic lapel mic (SHURE SM11)
- Playback and recording with PreSonus StudioLive RML32AI digital mixer
  - ~15dB difference between playback volume of foreground speech and background noise when measured near mic 1
  - 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 sec
  - 120 hours per microphone
- Volume of distant audio is lower than source audio
- Signal-to-noise ratio (SNR) of distant audio is much lower than source audio

Table 1: Average root mean square (RMS) level and SNR for source and distant audio

	Source Audio	Distant Audio
RMS Level (dbFS)	-19.97	-27.37
SNR (dB)	41.72	20.84

- SNR varies depending on noise conditions

Table 2: Average SNR for different background noise conditions

No Noise Added	Music	Television	Babble Speech
23.58 dB	19.18 dB	22.22 dB	18.40 dB

How to Obtain the Corpus

- Information and the download link on the corpus website: **voices.lab41.org**
  - Includes force aligned transcripts from LibriSpeech
  - Audio: 48kHz 24-bit WAV & 16kHz 16-bit WAV

Conclusions and Future Work

- Large free speech corpus with realistic recording conditions
- Several controlled variables:
  - Room conditions, mic distance, background noise
  - Angle between foreground loudspeaker and microphones
- Intended research areas:
  - Automatic speech recognition, speech activity detection
  - Speaker ID
  - Speech enhancement
  - Source separation
  - Event and background detection
  - Source distance and sound localization
- Similar collection planned in May-June 2018
  - 2 additional rooms with different acoustics
  - Additional mics to capture audio with lower SNR