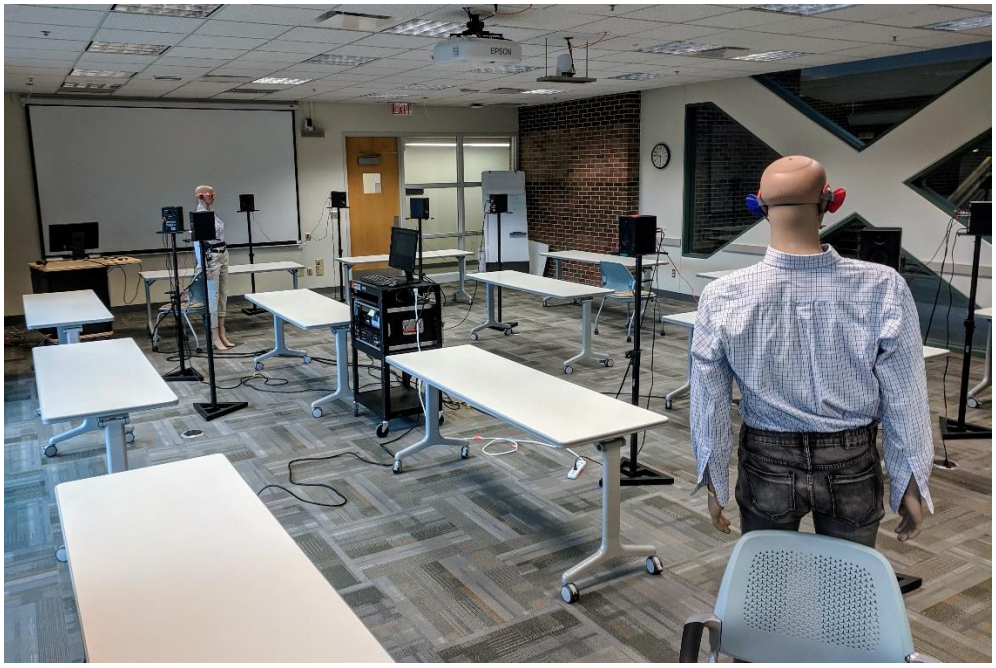
Massive Distributed Microphone Array Dataset

An open dataset of speech recorded with 160 microphones

Ryan M. Corey, Matthew D. Skarha, and Andrew C. Singer

University of Illinois at Urbana-Champaign

October 2019



AUDIO AT ILLINOIS

1 INTRODUCTION

Large, distributed microphone arrays could offer dramatic advantages for audio source separation, spatial audio capture, and human and machine listening applications. This dataset contains acoustic measurements and speech recordings from 10 loudspeakers and 160 microphones spread throughout a large, reverberant conference room.

The distributed microphone system contains two types of array: wearable microphone arrays of 16 sensors each placed near the ears and across the upper body, and tabletop arrays of 8 microphones each in enclosures designed to resemble voice-assistant speakers.

The dataset contains four types of audio data for each array device:

1. Exponential frequency sweeps from each of the ten loudspeakers, which can be used to calculate acoustic impulse responses and transfer functions,
2. Speech clips played separately through each loudspeaker, which can be used as ground-truth signals to simulate speech mixtures and evaluate source separation algorithms,
3. The same speech clips played simultaneously to simulate a cocktail party scenario, and
4. Recordings of ambient noise in the conference room with no loudspeakers active.

1.1 SOURCE DATA

The speech signals used in this dataset are derived from the University of Edinburgh Centre for Speech Technology Research Voice Cloning Toolkit (VCTK) corpus:

> Veaux, Christophe; Yamagishi, Junichi; MacDonald, Kirsten. (2017). CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/1994.

1.2 DATASET AVAILABILITY

This dataset is freely available on the Illinois Data Bank under a Creative Commons Attribution 4.0 International License CC BY, which requires attribution. Please cite this dataset as:

> Ryan M. Corey, Matthew D. Skarha, and Andrew C. Singer, *Massive Distributed Microphone Array Dataset*, University of Illinois at Urbana-Champaign, 2019. https://doi.org/10.13012/B2IDB-6216881_V1.

2 EXPERIMENTAL SETUP

2.1 TEST SIGNALS

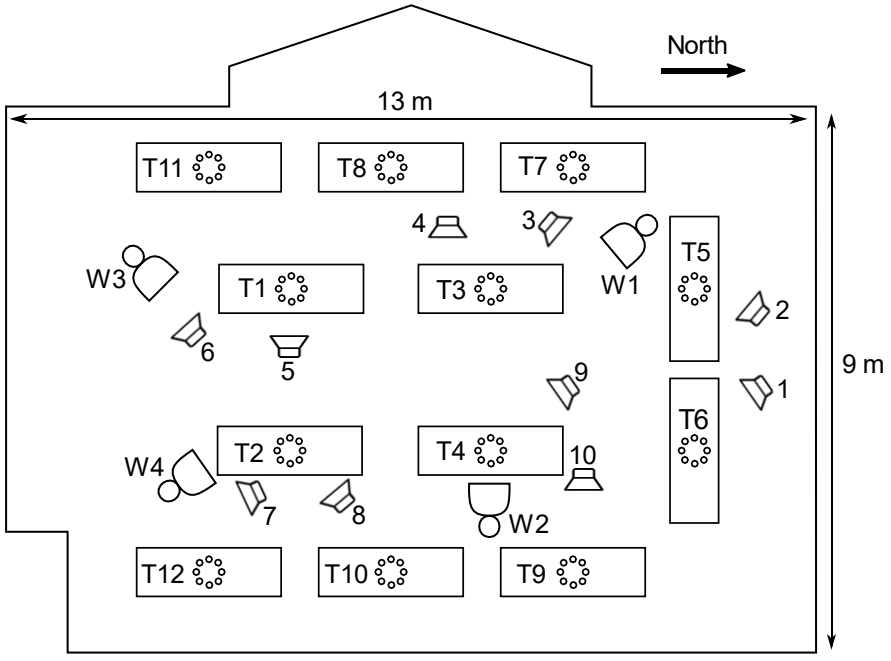
The chirp signals are identical exponential frequency sweeps with duration 20 seconds.

The 60-second speech signals used in this simulated cocktail party experiment were derived from the VCTK corpus of speech recorded in a hemi-anechoic chamber. For each of ten talkers, tabulated below, several sentences were concatenated together to resemble continuous speech. The genders and accents below are those reported in the VCTK corpus documentation.

Channel	VCTK Subject	Channel	VCTK Subject
1	245 (M, Irish)	6	239 (F, English)
2	253 (F, Welsh)	7	254 (M, English)
3	272 (M, Scottish)	8	258 (M, English)
4	333 (F, American)	9	288 (F, Irish)
5	363 (M, Scottish)	10	308 (F, American)

2.2 ROOM LAYOUT

Microphone array devices and loudspeakers were distributed throughout a large, reverberant conference room ($T_{60} \approx 800$ ms). The layout and numbering of the devices is shown below. Arrays W1–W4 are wearable arrays on mannequins and arrays T1–T12 are tabletop arrays in speaker-like enclosures.



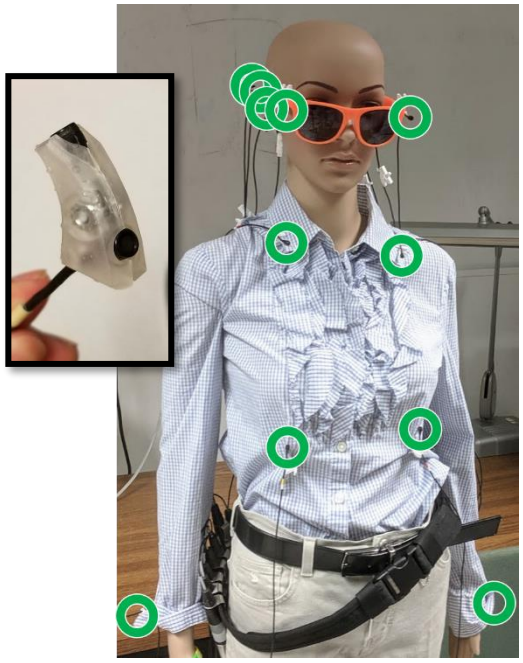
2.3 WEARABLE ARRAYS

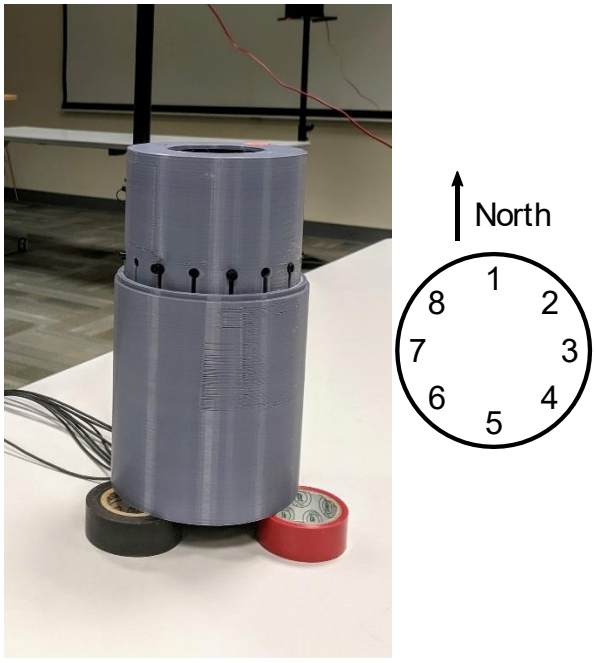
Sixteen microphones were attached to the clothing and accessories of plastic mannequins, including behind-the-ear earpieces that resemble hearing aid shells. Two plastic mannequins were each recorded in two locations, for a total of four wearable array locations. The microphone positions are summarized in the table below.

Channel	Location	Channel	Location
1	Left ear canal	9	Left lapel
2	Right ear canal	10	Right lapel
3	Left earpiece (front)	11	Left torso
4	Right earpiece (front)	12	Right torso
5	Left earpiece (rear)	13	Left back
6	Right earpiece (rear)	14	Right back
7	Left eyeglasses	15	Left cuff
8	Right eyeglasses	16	Right cuff

2.4 TABLETOP ARRAYS

The tabletop arrays contain eight microphones in plastic enclosures designed to resemble voice-assistant speakers. The microphone slots are arranged in a circular pattern with diameter 10 cm. The channels are numbered clockwise as viewed from above, with channel 1 facing North. The two enclosures were each recorded in six locations, for a total of twelve tabletop arrays.





2.5 Recording procedure

Recordings were captured by a 16-input, 10-output digital audio interface (Focusrite Scarlett 18i20 and OctoPre) at 24 bits and 48 kHz. The sensors are 16 Countryman B3 omnidirectional lavalier microphones with hard-wired XLR connections. The loudspeakers are 10 Presonus Eris E3.5 two-way studio monitors on stands 137 cm tall.

Because the equipment can only support 16 inputs at a time, the experiment was repeated 10 times and the microphones were moved between experiments. Wearable arrays were measured individually and tabletop arrays were measured in pairs (1 and 2, 3 and 4, etc.). Care was taken to ensure that the tables and loudspeakers did not move between recordings. The speech components of the recordings should therefore be consistent between arrays, but the background noise components are different.

3 Dataset Organization

Audio recordings are provided as multichannel WAVE files sampled at 48 kHz. In the table below, XX refers to the array number (01–04 for the wearable arrays and 01–12 for the tabletop arrays) and YY refers to the loudspeaker and talker number (01–10).

Filename	Length	Channels	Description
wearableXX/wearableXX_chirpYY.wav	21 sec	16	Chirp recording
wearableXX/wearableXX_speechYY.wav	61 sec	16	Individual speech
wearableXX/wearableXX_mix.wav	61 sec	16	Speech mixture
wearableXX/wearableXX_noise.wav	55 sec	16	Ambient noise
tabletopXX/tabletopXX_chirpYY.wav	21 sec	8	Chirp recording
tabletopXX/tabletopXX_speechYY.wav	61 sec	8	Individual speech
tabletopXX/tabletopXX_mix.wav	61 sec	8	Speech mixture
tabletopXX/tabletopXX_noise.wav	55 sec	8	Ambient noise
sources/chirp.wav	20 sec	1	Test chirp
sources/speechYY.wav	60 sec	1	Speech sample