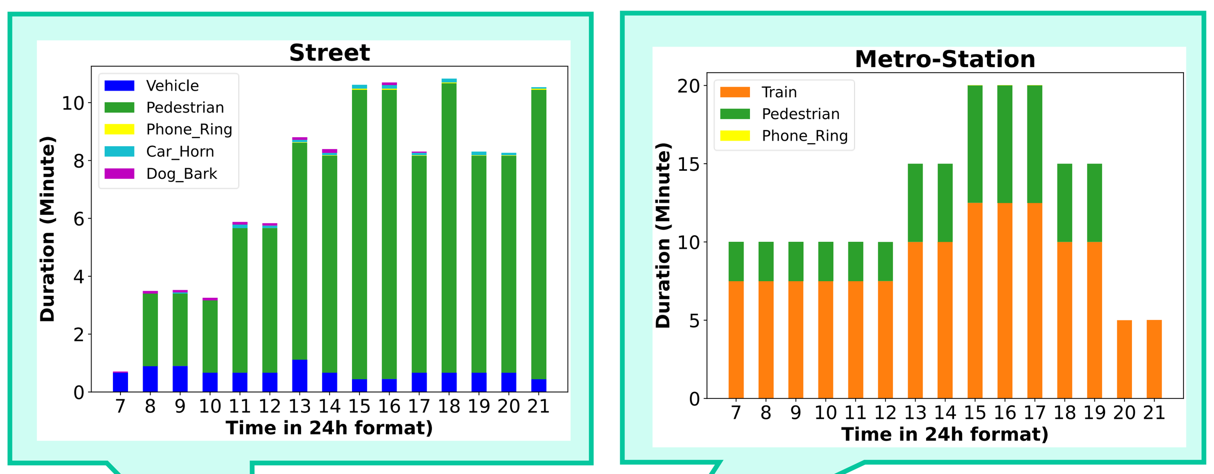
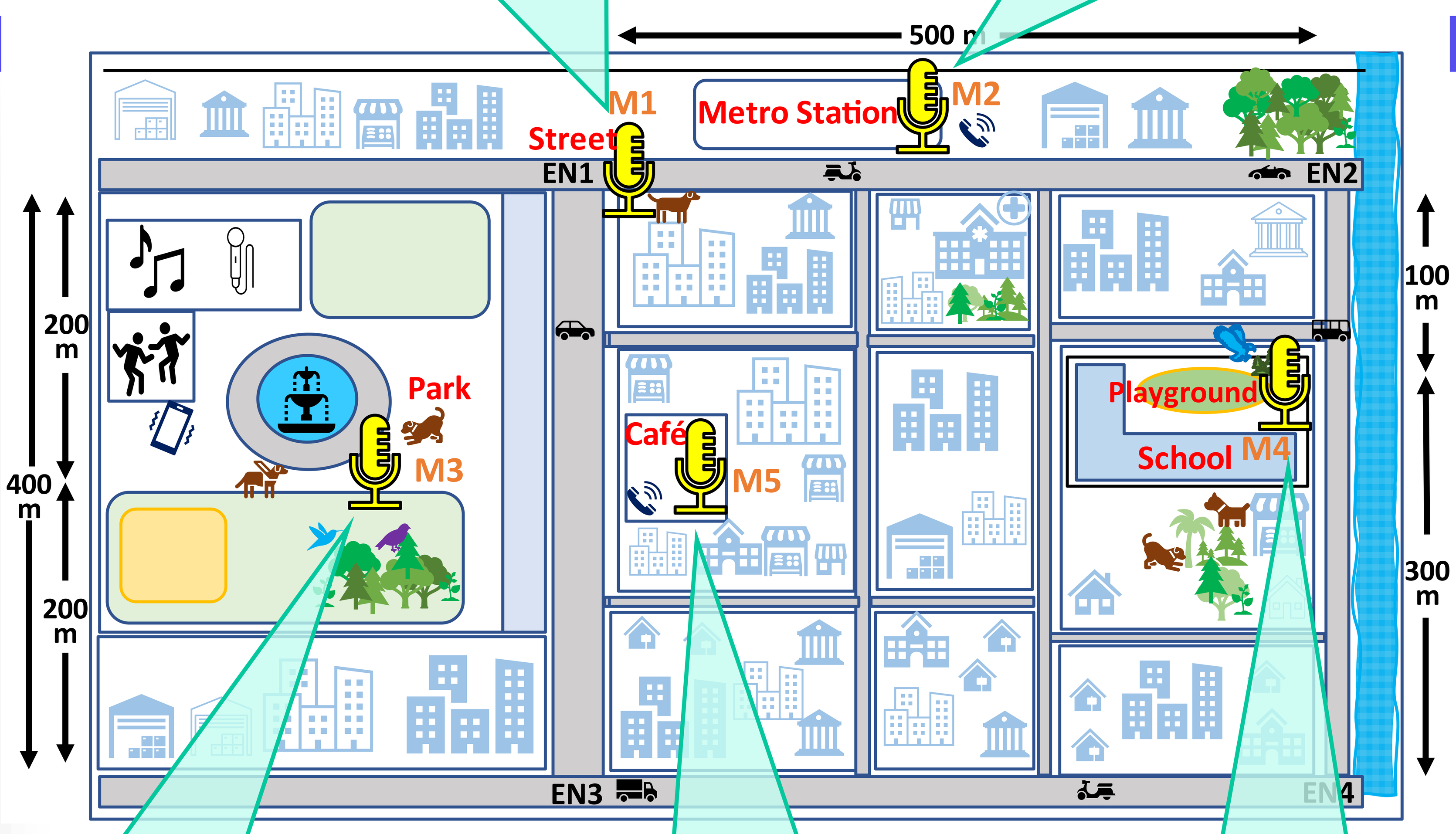
STeLiN-US: A Spatio-Temporally Linked Neighborhood Urban Sound Database

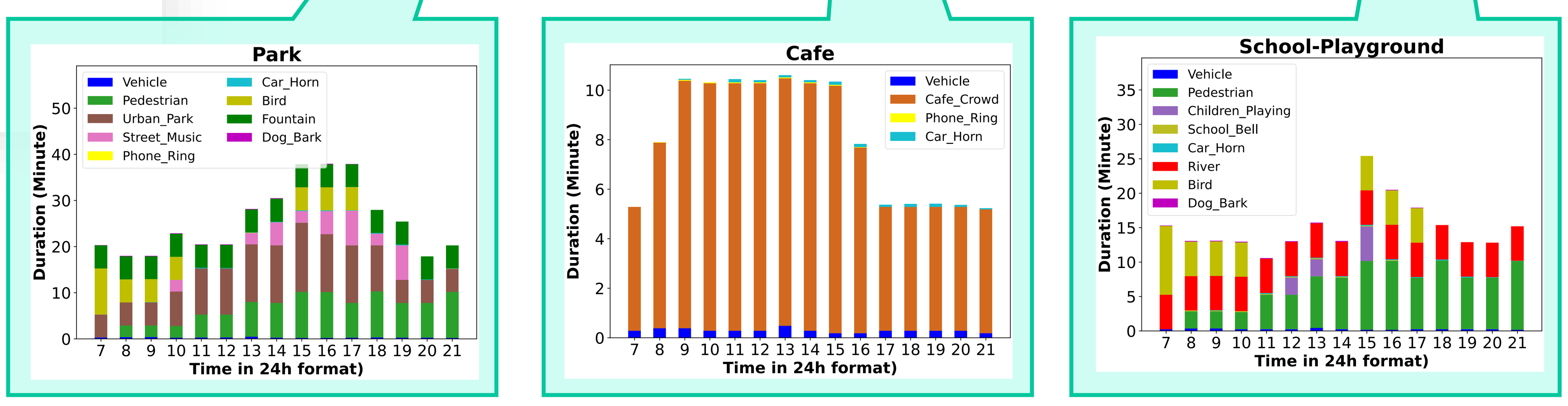
Behavioral-Informatics & Interaction-Computation Lab

Snehit Chunarkar, Bo-Hao Su, Chi-Chun Lee
Department of Electrical Engineering,
National Tsing Hua University, Taiwan

NTHU







Introduction:

- The dataset is semi-synthesized i.e., each sample is generated by leveraging diverse sets of real urban sounds with crawled information of real-world user behaviors over time.
- Proposed method helps create a realistic large-scale dataset, and we further evaluate it through perceptual listening tests.
- This neighborhood-based data generation opens up novel opportunities to advance user-centered applications with automated acoustic understanding.

Methodology:

Preconditions:

1. Design a Map to help visualize interconnections.
2. Map out microphone Locations.
3. Choose scene specific sound classes.
4. Follow real world pattern of surrounding sounds.

14 Classes: 8 Events & 6 Backgrounds

SONYC's [1] study & Google map popular time

Traffic Synthesis:

1. Design pattern for vehicle track.
2. Map out the distance of microphones from each other.
3. Use $Time = \frac{Distance}{Speed}$ to know the time of appearance of vehicle sound.

IDMT Traffic Dataset [2]

Scene Synthesis:

1. Distance Scale: Consider scaling factor based on how far the audio appears from the microphone.
2. Dense Scale: Add audio of same class based on level of dense environment aimed to create.
3. After both scaling, merge audio of each sound classes to synthesize the scene.

Discussion and Conclusion:

- Proposed STeLiN-US dataset [3] simulates the acoustic appearance of closely interconnected neighborhoods in urban areas.
- Accommodates the user-centered applications, e.g., If combined with ASR, it's performance can be analyzed based on surrounding.
- Incorporation of scene-specific events facilitates researchers in testing SED systems.
- The proposed synthesis approach can be dynamically scaled to model any environment.

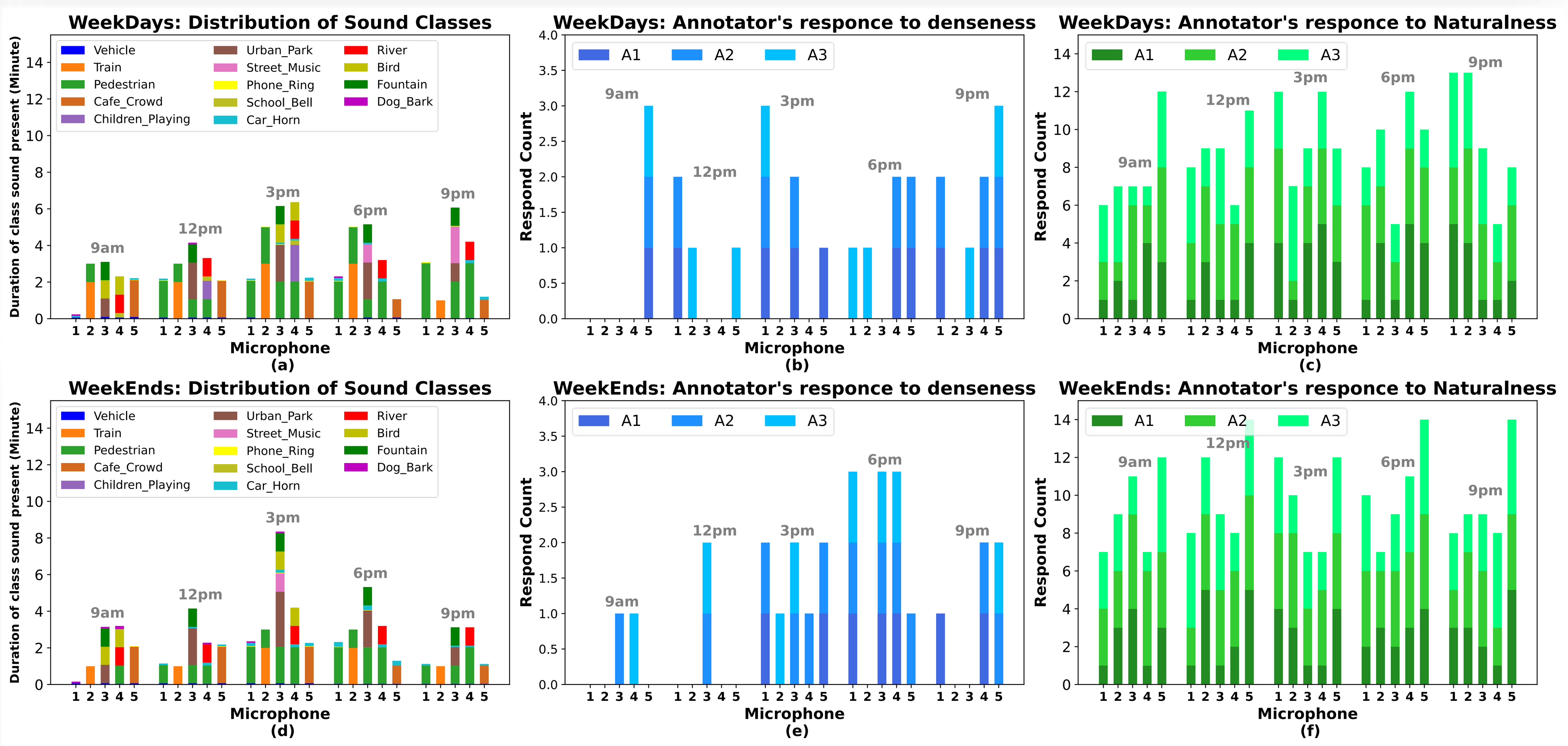
Summary:

1. Audio
   - Audio files: 525
   - Duration: 43hr 45min
2. Metadata
   - Strong annotations for events (8)
   - Information of vehicle sound file across microphones

Analysis by Listening Test:

- 50 audio samples (50 min) evenly distributed in all synthesized microphone locations and times selected for test.
- 2 Questions were asked:
  1. Do you think the sound is in rush hours (yes/no)?
  2. Rate for Naturalness of sound on scale 1-5?
- Unique 3 out of 6 annotator are selected for each audio annotation.

	Average	STD
Denseness	0.36	0.22
Naturalness	3.12	0.91



References:

1. M. Cartwright, A. E. M. Mendez, J. Cramer, V. Lostanlen, G. Dove, H.-H. Wu, J. Salamon, O. Nov, and J. Bello, "SONYC urban sound tagging (SONYC-UST): A multilabel dataset from an urban acoustic sensor network,"
2. Abeßer, S. Gourishetti, A. Katai, T. Clauß, P. Sharma, and J. Liebetrau, "Idmt-traffic: An open benchmark dataset for acoustic traffic monitoring research"
3. STeLiN-US: https://doi.org/10.5281/zenodo.8241539