Infrasonic Scene Fingerprinting for Authenticating Speaker Location

Kenji Aono*, Shantanu Chakrabartty† and Toshihiko Yamasaki‡
Department of Computer Science & Engineering* and Department of Electrical & Systems Engineering†, Washington University in St. Louis, St. Louis, MO, U.S.A.
Department of Information and Communications Engineering‡, The University of Tokyo, Tokyo, Japan

Introduction

This work explores a method for teasing out information on infrasonic cues that are beyond a standard microphone's frequency response by utilizing the acceleration vectors of cepstral features. Motivated by:

- Robust navigation cues exist in audio below 20 Hz
  - Nature utilizes this, state-of-the-art does not
- Geophysical phenomena affect a location's signature
- Passive method integrates into smartphones
- Audio allows for remote two-factor authentication

The results presented exceed performance reported in literature for audio-based localization schema.

Work	# of Locations	Active/ Passive	Error	Sample Length
[15]	5	Passive	20%	15s
[16]	10	Passive	80%	3600s
[17]	25	Active	15%	10s
[18]	33	Passive	31%	30s

Table 1: Published Audio-based Localization

Methods: Classifier

An overview of the scene classifier is shown below. 50% of collected data is used for training and 15% of the training data is used for cross validation.

- Input layer of 24 nodes (normalized)
- 69 hidden nodes (determined empirically)
- Standard w_0 bias node for stability
- Each hidden node is $\varphi = 2\left(1+\exp(-2n)\right)^{-1}$
- Output node is $\Sigma_M = \exp(n)\left(\Sigma_{\forall n}\exp(n)\right)^{-1}$

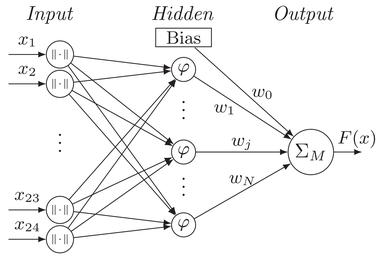

Figure 1: Diagram of proposed neural network.

Methods: Dataset

Data were collected from the campus of The University of Tokyo (UTokyo) and Michigan State University (MSU). UTokyo was collected on a 2013-era mid-tier smartphone with 44.1 kHz sampling rate and 16 bit depth. MSU data are recorded with a standalone recorder using 96 kHz sampling and 24 bit depth.

- Each sample of audio lasts 10 s
- Four samples per visit to each location
- Two microphone orientations per visit
- Two visits per location
  - UTokyo: 36 hours between visits
  - MSU: 1 week between visits
- Thus, a total of 80 s of audio per location

Figure 2: Example of locations under consideration.

Methods: Features

Evaluated hundreds of features and window parameters by utilizing Sequential Floating Forward Selection and a k-Nearest Neighbor classifier to find discriminative features. With consideration for ease-of-compute, the selected feature vector was:

- Low-pass filter with cutoff at 16 kHz
- 125 ms rectangular window with 50% overlap
- 10 dimension MFCC
- 0th cepstral and log energy
- $\Delta-\Delta$ (acceleration) features

Findings & Results

A synthetic input signal is used to validate our claim that features vary in presence of infrasonics. In the mixed case, 1 Hz tone is attenuated by 20 dB.

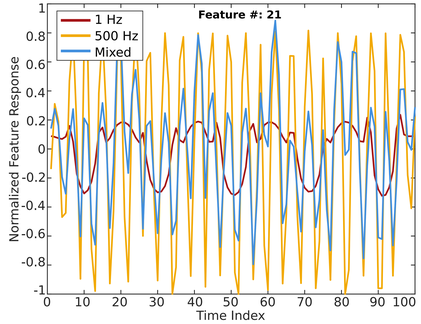

Figure 3: Infrasound only (1 Hz, magenta), audio-only (500 Hz, cyan), and superposition (orange).

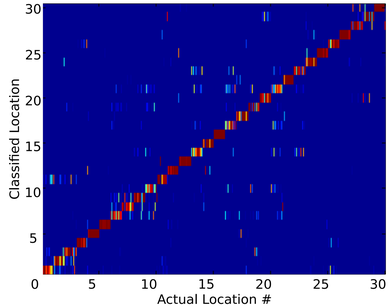

Figure 4: Confusion matrix for UTokyo (30) training.

Overall Classification Results

Results for both datasets collected during this work, as well as compared to the publically-available dataset from [18] show robust performance across a multitude of locations and recording aparatus. Of particular interest is the good performance shown when the audio is collected without the HVAC system running.

Dataset (# Scenes)	Test Error	Train Error	Train Time (s)
MSU[1] (15)	6.03%	5.83%	14.63
UTokyo (12)	4.11%	3.82%	10.71
UTokyo[2] (30)	9.17%	8.13%	52.62
Passive[3] (33) [18]	6.61%	5.59%	115.53
No HVAC (24) [18]	9.38%	7.65%	23.30
All[1,2,3] (78)	8.72%	7.35%	762.26

Table 2: Classification Error Rates (5-run Average)

- Trained on Intel Core i5-2520M
- Learned parameters passed to smartphone
  - Window parameters limit to 20 Hz update
  - Mid-tier can handle without issue
- Classifier shows low consecutive errors
- Median error rate under 10%
- Worst performing from UTokyo (30) is #21
  - Adjacent to locations #19, 24 and 27
- One location in [18] dataset with > 25% error
  - [18] reported 14 locations with error > 30%

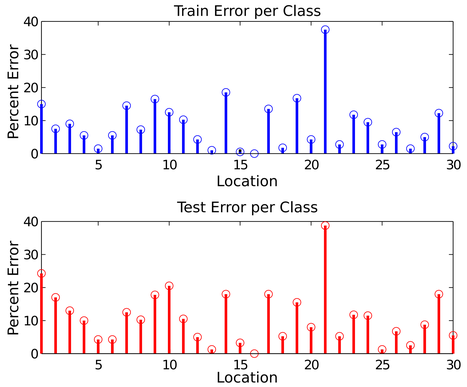

Figure 5: Per-location classification error in UTokyo (30).

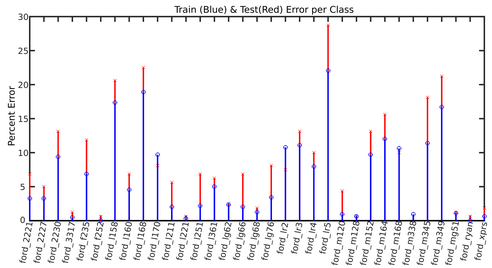

Figure 6: Per-location error for Passive (33) from [18].

Conclusion

We presented a method for scene classification of a speaker using ambient sounds captured on multiple platforms. The results suggest that acceleration-based cepstrals are mining infrasonic features that are robust across a variety of scenes and locations. Leveraging these features allows us to push the envelope of state-of-the-art with detection rates exceeding 90% for 78 locations. In a comprehensive platform, this audio modality would be augmented with information from visual sensors, GPS, WiFi, etc. to achieve peak performance.

See full paper for references. Lab websites: aimlab.seas.wustl.edu www.hal.t.u-tokyo.ac.jp

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. DGE-08022767 and DGE-1143954. K. Aono is an International Research Fellow of the Japan Society for the Promotion of Science (GR14001). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.