INFRASONIC SCENE FINGERPRINTING FOR AUTHENTICATING SPEAKER LOCATION

Kenji Aono[⋆] *Shantanu Chakrabartty*[†] *Toshihiko Yamasaki*[‡]

[⋆]Department of Computer Science & Engineering, [†]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

ABSTRACT

Ambient infrasound with frequency ranges well below 20 Hz is known to carry robust navigation cues that can be exploited to authenticate the location of a speaker. Unfortunately, many of the mobile devices like smartphones have been optimized to work in the human auditory range, thereby suppressing information in the infrasonic region. In this paper, we show that these ultra-low frequency cues can still be extracted from a standard smartphone recording by using acceleration-based cepstral features. To validate our claim, we have collected smartphone recordings from more than 30 different scenes and used the cues for scene fingerprinting. We report scene recognition rates in excess of 90% and a feature set analysis reveals the importance of the infrasonic signatures towards achieving the state-of-the-art recognition performance.

Index Terms— Acoustic Filtering, Classifier, Localization, Authentication, Infrasound

1. INTRODUCTION

Speech-based authentication has several advantages compared to other biometric techniques in that it can be performed remotely using a mobile device without the need for any specialized equipment [1]. However, the remote access capability also introduces vulnerabilities that may result in compromised security of the biometric technique [2].This vulnerability could be mitigated by using a two-stage authentication (as in passcodes) technique or by using side-channel information like the location of the acoustic channel or background of the speaker [3–5].

In this regard, infrasound – or sound with frequencies less than 20 Hz – has been known to capture signatures that can be used to identify many geophysical phenomena, such as seismic activity, air turbulence, and wind noise [6–8]. Some of these signatures are strongly correlated with the ambient acoustic environment and hence could be used to uniquely identify or fingerprint the spatial location. For instance, an indoor room can form a Helmholtz resonator which can sustain infrasonic standing waves, whose spatial properties depend on the room size and location of the windows or doors [9, 10]. Literature has shown that animals, such as the common pigeon, can detect infrasound around 1 Hz and that these cues may aid in their ability to navigate accurately across large spans where other senses such as smell and sight are limited in their usefulness, yet these features have not been leveraged in current state-of-the-art localization or authentication systems [6, 11–13]. Exploiting the audio modality for location tracking can provide an alternative to the likes of GPS, whilst maintaining a low-cost and low-burden to the user [14].

In this paper, we will explore how these infrasonic cues can be extracted using a standard smartphone. This presents a challenge because smartphone microphones are optimized for the human auditory range, often in the range of 100 Hz to several kHz. The hypothesis is that by measuring the acceleration of the auditory cepstral features, one could infer the underlying infrasound that is uniquely present in an acoustic scene. In addition, these infrasonic cues can provide features that are robust to sources typically considered "noise" (speech, music, and other higher frequency sources) for the purposes of scene fingerprinting and classification. We verify our claim by conducting a scene classification in a variety of settings, such as: laboratory, classroom, elevators, and multi-purpose rooms – a representative subset of which are shown in Fig. 1.

1.1. Related Works: Audio-based

The notion of using auditory features for localization has been studied in many forms by several researchers [15–18] to varying levels of success. In this paper, we will use the published results from [18] as the benchmark target since it achieved a high accuracy in passive sensing from a smartphone; moreover, their dataset has been released for public use. In passive sensing, the audio data are collected from the microphone without disturbing the environment, whereas in active sensing, a stimuli is generated and the environment's response to the stimuli is measured in an attempt to obtain more discriminating information, akin to the SONAR concept [17].

This material is based upon work supported by the National Science Foundation under Grant Nos. DGE-0802267 and DGE-1143954. K. Aono is an International Research Fellow of the Japan Society for the Promotion of Science (GR14001).

Fig. 1. Rooms under consideration from UTokyo dataset (See Section 2.2). Clockwise from bottom-left are location #s: 5, 15, 19, and 27.

Researchers have shown that an active schema can generate enough information that a room's geometry can be reconstructed, albeit under specific recording parameters [19, 20]. Table 1 shows classification results that have been published based on audio features alone. It should be noted that the audio features could be used in conjunction with the aforementioned modalities to provide superior location estimates.

Work	# of Locations	Active/ Passive	Accuracy	Sample Length
[15]	5	Passive	80%	15s
[16]	10	Passive	20%	3600s
[17]	25	Active	85%	10s
[18]	33	Passive	69%	30s

Table 1. Contemporary Audio-based Localization

2. PROPOSED METHOD

2.1. Overview

One design objective for the design of an acoustic scene recognition platform is to have a path towards a low-power implementation that can be deployed on a smartphone. Therefore, we have used a filter bank for feature extraction and a simple artificial neural network (ANN) classifier, both of which can be fabricated on dedicated silicon to operate with sub-microwatt power consumption [21–23]. The ANN backend approach will also allow for any computationally intensive tasks to be delegated to a remote server, while allowing the client side to maintain a low-power profile during operation.

2.2. Datasets

Testing and validation is conducted on three different sets of data. The first (henceforth referred to as: UTokyo dataset) was collected using a mid-tier Android smartphone released in 2013, the LG Optimus L-05E with a sampling rate of 44.1 kHz and 16 bit depth; it contains a sample of 12 rooms from a single building as well as a collection of 30 locations (including the aforementioned 12 rooms) across multiple buildings – all samples are collected from the Hongo main campus of The University of Tokyo, Japan. Data were collected during two sessions with a temporal spacing of approximately 36 hours between sessions. Each session collected four samples of audio lasting 10 s per sample. Of the four samples, two were taken near what would be considered the entrance of the room with a vertical and horizontal smartphone orientation. The remaining two samples were collected near the center of the room, also with a varying smartphone orientation. In the case of locations that did not have an obvious entrance (such as hallways or outdoors), the two sets of samples were chosen with at least 6 m of spacing. To clarify, each location will have $2 \text{ visits} \cdot 4 \text{ samples} \cdot 10 \text{ s} = 80 \text{ s}$ of data.

The second dataset (MSU dataset) was recorded with an Olympus Linear PCM Recorder LS-10 at a sampling rate of 96 kHz and 24 bit depth, and contains 15 scenes of audio, collected from within the Michigan State University (MSU) Engineering Building. A similar procedure as in the UTokyo dataset for location/orientation was carried out. In this dataset, the visits were timed to be one week apart instead of one-and-a-half days apart.

The final dataset being considered was from previously published work by Tarzia *et al.*, we utilize the 33 room passive and 24 room with the HVAC off subsets, for details on these datasets, we refer readers to [18, 24].

2.3. Feature Selection

For decades, the Mel-frequency cepstral coefficients (MFCCs) have been a popular transformation for purposes of speech recognition and have also proven useful in modeling urban soundscapes [25, 26]. It is common knowledge in the field that features such as the log energy, zero crossing rate, energy envelope, spectral power, delta coefficients, and wavelet decompositions may provide features with favorable characteristics when applied to audio-based classifiers. Based on this knowledge, and after considering a variety of window sizes ranging from several ms to hundreds of ms, as well as their shape and overlap, we created an initial feature vector dimension in excess of 600 for each sample in the passive dataset from [18].

To determine the features that are conducive to scene fingerprinting, we apply a Sequential Floating Forward Selection (SFFS) algorithm to rank the features. For the purposes of this paper, we did not fully train an ANN classifier during each iteration, instead, we used a simple k-Nearest Neighbor classifier to save time during SFFS – even then this ranking took over 600 hours of computation. All top ten of the ranked features were based on MFCC $\Delta-\Delta$, a promising result that shows the importance of cepstral acceleration features. In examining the top 100 ranked features (out of 578) all 100% of the $\Delta-\Delta$ and 0^{th} cepstral features tested were selected, re-

gardless of the parameters used, and approximately 50% of the tested energy and mode features ranked in the top 100. The remaining features had less than 20% of their features ranked in the top 100.

From the SFFS results, the most discriminative set of 10 MFCCs with a sub-16 kHz frequency response were selected, with the parameters of 50% overlap and 125 ms rectangular time window. The 0^{th} cepstral and log energy of the sample were also selected for incorporation into the feature set, thereby resulting in a feature vector with a dimension of 12; from this, we calculated the $\Delta-\Delta$ features to get a final feature set with a dimension of 24.

2.4. Classifier

Scene classification is done by a basic two-layer ANN from MATLAB with training being done offline from the smartphone. In this manner, the pre-calculated biases and weights can be uploaded to the smartphone to allow for a quick and low-power classification in the field.

To match the feature set dimension, the input layer of the ANN has 24 nodes and accepts a feature vector that is normalized to a magnitude of one. The optimal number of hidden nodes was empirically determined as 69, with an additional w_0 bias node to promote stability during the training process. Each node uses a hyperbolic tangent sigmoid activation that is characterized as $\varphi = 2\left(1+\exp(-2n)\right)^{-1}$. The output node uses a SoftMax function $\Sigma_M = \exp(n)\left(\Sigma_{\forall n}\exp(n)\right)^{-1}$.

MATLAB's built-in cross entropy fitness function is utilized during the training process, hence resulting in a smoother network function with improved generalization. 50% of the collected data was used for training, with 15% of that data being used as a validation set to reduce the chance of overfitting the network.

3. FINDINGS & RESULTS

3.1. Synthetic Infrasound

To validate the claim that acceleration-based cepstral features can provide discriminatory features in the presence of infrasonic signatures, an ideal single frequency sinusoidal of 1 Hz at a sampling rate of 44.1 kHz was generated and passed through a filterbank with parameters based on the SFFS ranking. The corresponding feature vector was non-zero, thereby confirming that features were being generated based only on an infrasonic stimuli. Subsequently, a 500 Hz sinusoidal is also tested to confirm the feature extraction method also works in more traditional audio ranges. Finally, a superposition of the 500 Hz with the 1 Hz (with 20 dB attenuation applied to the lower frequency to emulate the frequency response of a typical microphone) signal is also tested. The 21^{st} dimension of the resulting feature vector, which is a $\Delta-\Delta$ or acceleration feature, is presented in Fig. 2. From inspection, it is clear that the proposed feature extraction will generate differing feature vectors when presented with infrasonic cues, even if said infrasonics are mixed with higher frequency sources.

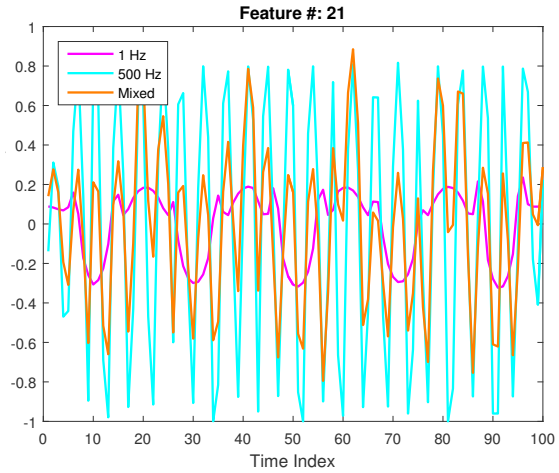

Fig. 2. Plots showing the selected feature set is discriminating between infrasound only (1 Hz in magenta), audio-only (500 Hz in cyan), and a mixture of the two (in orange).

3.2. Classification Results

After data are collected and the features extracted, 50% of the feature vector is used for training the classifier and the remaining 50% is relagated for testing the performance of the trained classifier. Although the training of the classifier could be done on-the-fly by a smartphone, it is more efficient to precalculate the ANN parameters ahead-of-time; therefore, the classifier training was completed on a 2011-era Core i5 dual core system (Intel(R) Core(TM) i5-2520M) with MATLAB R2015b. The resulting bias and weight values were sent to a smartphone to carry out the final classification. Since the implementation of training code takes advantage of parallel computing principles, it is possible to scale this method to an industrial-sized dataset and maintain reasonable training time by utilizing larger computing services. In practice, this batch training need only occur once, with continual updates coming from online training methods performed by the smartphone to maintain optimal classifier performance.

After training, the ANN is fed the extracted features during an epoch and will return a maximally responding node as the classified location. By virtue of the selected window size and overlap, the ANN can handle 20 updates per second, a rate that would be high enough for real-time tracking or authentication purposes. To demonstrate the classification error of the system, we supply a trained ANN with the extracted features from the UTokyo 30 location dataset. Fig. 3 shows the classification error per location for the training and testing datasets. The majority of scenes have an error rate below 20%, and the average error rate remains below 10% – only a single location suffered bad detection rates (which is defined as $100\% - \text{error rate}$). This particular scene is an

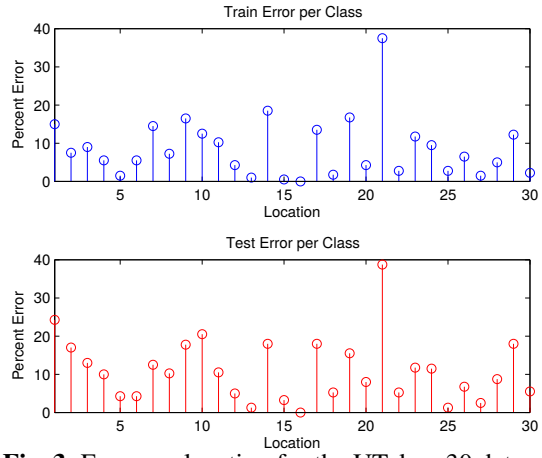

Fig. 3. Error per location for the UTokyo 30 dataset.

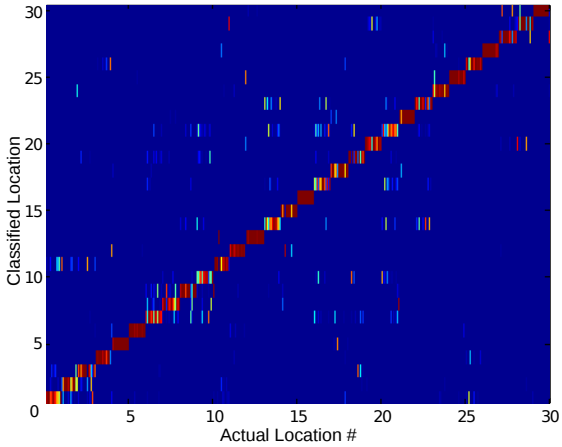

Fig. 4. Confusion matrix of artificial neural network on UTokyo 30 dataset.

open hallway located near a bank of elevators (elevator's interior is a separate scene under test), it is not evident why the extracted features from this location consistently exhibit poor classification performance. The five-run average detection rate for the UTokyo 30 training and testing datasets were 91.87% and 90.83%, respectively. The resulting confusion matrix is presented as Fig. 4, with a blue entry representing a -1, or deactivated neuron and a red corresponding to a $+1$ or strongly responding neuron. From this confusion matrix, we can see that as the input location is varied during testing (sequentially from 1 to 30), the scene classification properly tracks the ground truth. We can also note that when the classification is incorrect, it tends to be a spurious event and not one that is sustained.

Presented in Table 2 are the overall classification errors, averaged over five separate runs for multiple datasets. When compared to the dataset from [18], the proposed method shows an improvement of around 25 percentage points with respect to detection rates (results from [18] at 69%, our proposed method 94%). We are also reporting low error rates for the MSU dataset and both the UTokyo 12 dataset which was limited to a single floor on a particular building and the UTokyo 30 dataset spanning multiple buildings. Another finding is

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

Table 2. Classification Error Rates (five-run average)

that the proposed method has a relatively low classification error even when combining multiple datasets recorded on varying devices as we obtained a 7.35% error when classifying among 78 locations (MSU(15)+UTokyo(30)+Passive(33)). Of particular interest is the strong performance of this method in the no HVAC subset from [18], with a detection rate in excess of 90% even without the fans or compressors from the HVAC contributing to the signature, which suggests other uniquely identifying sound generators are present in the environment. Considering that audio attenuates proportional to the square of its frequency (i.e. low frequency sounds tend to pass through barriers more freely and with less absorption or reflection), and the lack of speech or other common sources of higher frequency, it is very well possible that infrasound from distant locations were acting as a triangulating beacon and contributing to the unique fingerprinting of the scenes.

4. CONCLUSION & REMARKS

We presented a method for scene classification of a speaker using ambient sounds captured on a smartphone as the source modality. These results suggest that the use of acceleration-based cepstral features are mining infrasonic features that are robust across a variety of scenes and locations. By leveraging these features, we are maintaining a high degree of detection even on large dataset, previously unmatched in the literature [15–18]. With detection rates exceeding 90% for datasets containing several dozen scenes, it is now possible to say that these features alone would suffice for a location authentication service — even more so when fused with traditional methods prevalent in industry. The audio modality can also supplant the traditional methods (GPS, WiFi, etc.) when they fail, thus providing end users with a high accuracy method for determining their location in more situations. A facet that still requires further investigation is the stability of these features as time progresses. Although the UTokyo and MSU datasets had one-and-a-half days and one week between sample measurements, respectively, we did not have access to audio data over extended periods of time. On a seasonal scale, it is possible that switching from air-conditioning to heating could yield significantly different features and impact the ambient acoustics of a location.

5. REFERENCES

[1] J. P. Campbell Jr, "Speaker Recognition: A Tutorial," *Proc. IEEE*, vol. 85, no. 9, Sep 1997, pp. 1437–1462.

[2] J. H. Lee and R. M. Buehrer, "Characterization and detection of location spoofing attacks," *Journal of Communications and Networks*, vol. 14, no. 4, 2012, pp. 396–409.

[3] Y. Kawamoto, *et al.*, "Effectively Collecting Data for the Location-Based Authentication in Internet of Things," *IEEE Syst. J.*, Pre-Print, pp. 1–9.

[4] S. Billeb, *et al.*, "Efficient Two-stage Speaker Identification based on Universal Background Models," *Biometric Special Interest Group (BIOSIG), 2014 International Conference of the*, Darmstadt, 2014, pp. 1–6.

[5] L. Xiao, *et al.*, "Fingerprints in the Ether: Channel-Based Authentication," *Securing Wireless Communications at the Physical Layer*, Springer US, 2010, pp. 311–333.

[6] J. T. Hagstrum, "Infrasound And The Avian Navigational map," *The J. of Experimental Biology*, vol. 203, Mar. 2000, pp. 1103–1111.

[7] J. A. Nystuen and H. D. Selsor, "Weather Classification Using Passive Acoustic Drifters," *J. of Atmospheric and Oceanic Technology*, vol. 14, 1996, pp. 656–666.

[8] U. Fehr, "Measurements of Infrasound from Artificial and Natural Sources," *J. of Geophysical Research*, vol. 72, no. 9, May 1967, pp. 2403–2417.

[9] W. T. Plummer, "Infrasonic Resonances in Natural Underground Cavities," *J. Acoust. Soc. Am.*, vol. 46, no. 5, 1969, pp. 1074–1080.

[10] D. Olivia, *et al.*, "Measurements of low frequency noise in rooms," *Finnish Institute of Occupational Health*, Helsinki, Finland, 2011.

[11] H. Zhao and H. Malik, "Audio Recording Location Identification Using Acoustic Environment Signature," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 11, 2013, pp. 1746–1759.

[12] M. Fan *et al.*, "Public Restroom Detection on Mobile Phone via Active Probing," *ACM ISWC'14*, Seattle, WA, U.S.A., Sep 13–17, 2014.

[13] M. L. Kreithen and D. B. Quine, "Infrasound Detection by the Homing Pigeon: A Behavioral Audiogram," *J. Comp. Physiol.*, vol. 129, Aug. 1978, pp. 1–4.

[14] K. Muthukrishnan *et al.*, "Towards Smart Surroundings: Enabling Techniques and Technologies for Localization," *First Int. Workshop Loc. Content-Awareness (LOCA)*, 2005.

[15] J. Du *et al.*, "Catch You as I Can: Indoor Localization via Ambient Sound Signature and Human Behavior," *Int. J. Dist. Sensor Networks,* vol 2013, id. 434301, pp. 1–16.

[16] M. Azizyan and R. R. Choudhury, "SurroundSense: Mobile Phone Localization Using Ambient Sound and Light," *ACM SIGMOBILE, Mobile Computing Communications Review*, vol. 1, no. 13, 2009, pp. 69–72.

[17] M. Rossi *et al.*, "RoomSense: An Indoor Positioning System for Smartphones using Active Sound Probing," *ACM AH'13*, Stuttgart, Germany, Mar 7–8, 2013.

[18] S. Tarzia *et al.*, "Indoor Localization without Infrastructure using the Acoustic Background Spectrum," *ACM MobiSys'11*, Bethesda, MD, U.S.A., Jun 28–Jul 1, 2011.

[19] P. Lazik and A. Rowe, "Indoor pseudo-ranging of mobile devices using ultrasonic chirps," *SenSys '12 Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*, Nov. 2012, pp. 99–112.

[20] I. Dokmanić *et al.*, "Acoustic echoes reveal room shape," *PNAS*, vol. 110, no. 30, Jul. 2013, pp. 12186–12191.

[21] J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress," *Neurocomputing*, vol 74, 2010, pp. 239–255.

[22] K. Aono *et al.*, "Exploiting Jump-Resonance Hysteresis in Silicon Auditory Front-Ends for Extracting Speaker Discriminative Formant Trajectories," *Biomedical Circuits and Systems, IEEE Transactions on*, vol. 7, no. 4, Aug 2013, pp. 389–400.

[23] ——, "Exploiting Jump-Resonance Hysteresis in Silicon Cochlea for Formant Trajectory Encoding," *Circuits and Systems (MWSCAS), 2012 IEEE 55th International Midwest Symposium on*, 2012, pp. 85–88.

[24] S. Tarzia, "Acoustic sensing of location and user presence on mobile computers," Ph.D. dissertation, Dept. Elec. Eng. and Comp. Sci, Northwestern Univ., Evanston, IL, 2011.

[25] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoustics, Speech and Signal Proc.*, vol. ASSP-28, no. 4, Aug. 1980, pp. 357–366.

[26] J. Aucouturier *et al.*, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *Acoustical Society of America*, vol. 122, no. 2, Aug. 2007, pp. 881–891.